VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STROJNÍHO INŽENÝRSTVÍ

FACULTY OF MECHANICAL ENGINEERING

ÚSTAV MATEMATIKY

INSTITUTE OF MATHEMATICS

VYUŽITÍ FUZZY MNOŽIN VE SHLUKOVÉ ANALÝZE SE ZAMĚŘENÍM NA METODU FUZZY C-MEANS CLUSTERING

FUZZY SETS USE IN CLUSTER ANALYSIS WITH A SPECIAL ATTENTION TO A FUZZY C-MEANS CLUSTERING METHOD

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. Assa Camara

VEDOUCÍ PRÁCE
SUPERVISOR

doc. RNDr. Libor Žák, Ph.D.

BRNO 2020

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA STROJNÍHO INŽENÝRSTVÍ

Zadání diplomové práce

Ústav:	Ústav matematiky
Studentka:	**Bc. Assa Camara**
Studijní program:	Aplikované vědy v inženýrství
Studijní obor:	Matematické inženýrství
Vedoucí práce:	**doc. RNDr. Libor Žák, Ph.D.**
Akademický rok:	2019/20

Ředitel ústavu Vám v souladu se zákonem č.111/1998 o vysokých školách a se Studijním a zkušebním řádem VUT v Brně určuje následující téma diplomové práce:

Využití fuzzy množin ve shlukové analýze se zaměřením na metodu Fuzzy C–means Clustering

Stručná charakteristika problematiky úkolu:

Shluková analýza je součástí matematické oblasti, která se zabývá hledáním možných závislostí v datech. V některých případech není vhodné použití klasických množin při hledání možných shluků. Je potřeba použít vágnější popis – např. pomocí fuzzy množin. Práce by se měla zabývat využitím fuzzy množin při shlukování a zvláště pak metodou Fuzzy C–means – definovanou J.C. Bezdeken.

Cíle diplomové práce:

Stručný popis shlukovacích metod.
Popis metod využívající fuzzy množiny.
Popis a použití metody Fuzzy C–means.
Aplikace shlukovacích metod (včetně Fuzzy C–means) na reálných datech.
Vyhodnocení úspěšnosti použití shlukovacích metod.

Seznam doporučené literatury:

LUKASOVÁ, A. a ŠARMANOVÁ, J. Metody shlukové analýzy, SNTL, Praha 1985, ISBN 04-014-85.

BEZDEK, J. C. Pattern Recognition with Fuzzy Objective Function Alghorithms. Plenium Press, New York 1981, ISBN 978-1-4757-0452-5.

DUBOIS, D. a PRADE, H. Fuzzy Sets and Systems: Theory and Applications, Academic Press, New York, 1980, ISBN 0–12–222750–6.

DUNN, J. C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters". Journal of Cybernetics. 3 (3): 32–57. ISSN 0022-0280.

Termín odevzdání diplomové práce je stanoven časovým plánem akademického roku 2019/20

V Brně, dne

L. S.

prof. RNDr. Josef Šlapal, CSc. ředitel ústavu	doc. Ing. Jaroslav Katolický, Ph.D. děkan fakulty

Abstrakt

Táto práca sa zaoberá zhlukovou analýzou, a podrobnejšie zhlukovacími metódami, ktoré používajú fuzzy množiny. V teoretickej časti sú popísané zhlukovacie metódy a transformácie potrebné na zhlukovú analýzu. V praktickej časti aplikujeme na reálne dáta. Tieto dáta predstavujú vstupné dáta z chemicko-transportného modelu CMAQ, ktorý sa používa na získanie výpočtu koncentrácii znečisťujúcich látok v atmosfére. Na tieto dáta aplikujeme dve rôzne metódy, metódu k-means a fuzzy c-means. Pre metódu fuzzy c-means porovnáva dva rôzne prístupy k zvoleniu optimálneho váhového exponentu. Porovnali sme takto vytvorené 3 zhlukovacie štruktúry. Výsledné zhluky si boli podobné a však metóda fuzzy c- means s vyššiu hodnotou váhového exponentu vytvorila zhluky, ktoré nemali žiadnu podobnosť so zhlukovanými veličinami. V závere sme vytvorili regresný model na nájdenie vzťahu medzi vstupnými a výstupnými dátami modelu CMAQ.

kľúčové slová

Zhluková analýza, zhlukovanie fuzzy c-means, váhový exponent, zhlukovanie k-means, CMAQ, kvalita ovzdušia.

Abstract

This master thesis deals with cluster analysis, more specifically with clustering methods that use fuzzy sets. Basic clustering algorithms and necessary multivariate transformations are described in the first chapter. In the practical part, which is in the third chapter we apply fuzzy c-means clustering and k-means clustering on real data. Data used for clustering are the inputs of chemical transport model CMAQ. Model CMAQ is used to approximate concentration of air pollutants in the atmosphere. To the data we will apply two different clustering methods. We have used two different methods to select optimal weighting exponent to find data structure in our data. We have compared all 3 created data structures. The structures resembled each other but with fuzzy c-means clustering, one of the clusters did not resemble any of the clustering inputs. The end of the third chapter is dedicated to an attempt to find a regression model that finds the relationship between inputs and outputs of model CMAQ.

keywords

Cluster analysis, fuzzy c-means clustering, weighing exponent, k-means clustering, CMAQ, air quality.

CAMARA, Assa.*Využití fuzzy množin ve shlukové analýze se zaměřením na metodu Fuzzy C-means Clustering* . Brno, 2020 [cit. 2020-01-07]. 49s Dostupné z: https://www.vutbr.cz/studenti/zav-prace/detail/121466. Diplomová práce. Vysoké učení technické v Brně, Fakulta strojního inženýrství, Ústav matematiky. Vedoucí práce doc. RNDr. Libor Žák, Ph.D.

Prehlasujem, že som diplomovú prácu *Využití fuzzy množin ve shlukové analýze se zaměřením na metodu Fuzzy C-means Clustering* vypracovala samostatne pod vedením doc. RNDr. Libor Žák, Ph.D. s použitím materiálov uvedených v zozname literatúry.

Assa Camara

Chcela by som sa poďakovať Jane Matejovičovej zo Slovenského hydrometeorolického ústavu, za čas a rady, ktoré mi venovala. Mojej sestre Anne, priateľovi Lukášovi a mame Jane za všetku mentálnu oporu aj pomoc pri oprave textu. V neposlednej rade by som chcela poďakovať vedúcemu tejto diplomovej práce doc. Liborovi Žákovi za jeho odborné vedenie.

Assa Camara

Obsah

Úvod

Zhluková analýza je súčasťou nekontrolovaného učenia a časťou nepriameho data-miningu. Našla využitie v mnohých oblastiach ľudského života od biológie, cez počítačové vedy a rôzne technické oblasti po spoločenské vedy. Aj napriek tomu, že vo vedeckej praxi je zhluková analýza používaná a skúmaná aspoň od roku 1939, vďaka technologickému pokroku v priebehu posledných desaťročí sa jej aplikačné možnosti neustále rozširujú a je stále predmetom vedeckého bádania.

Fuzzy množiny, alebo aj vágne množiny, majú schopnosť zachytiť nepresnosti skutočného života lepšie ako mnohé striktné matematické pojmy. Vágne pojmy a vágne množiny sú predsa v živote typické, červené jablká v obchode nie sú úplne červené. Ľudský život je sprevádzaný množstvom vágnych pomenovaní a aplikovaná matematika sa snaží zachytiť túto vágnosť aj použitím fuzzy množín. V tejto práci sa budeme venovať obom pojmom zhlukovania a fuzzy množín a ich aplikácii.

V prvej kapitole sa budeme venovať teórii potrebnej pre zhlukovú analýzu a položíme teoretické základy hierarchického zhlukovania, nehierarchického zhlukovania a zhlukovania neurónovými sieťami. Detailnejšie popíšeme zhlukovacie algoritmy, na ktorých sú založené metódy používajúce fuzzy množiny a ich fuzzy logiky a následne popíšeme aj tieto metódy samotné. Popíšeme základy regresných modelov, ktoré použijeme na aproximáciu koncentrácii v závere tretej kapitoly.

Zhlukovú analýzu aplikujeme na dáta kvality ovzdušia, ktoré nám poskytli pracovníci Úseku kvality ovzdušia zo Slovenského hydrometeorologického ústavu (SHMU). Kvalita ovzdušia úzko súvisí s ochranou ľudského zdravia a aj životného prostredia. V priebehu minulého storočia sa pozornosť vedeckej aj laickej verejnosti upriamila na niekoľko environmentálnych problémov, ako sú kyslé dažde či smogové epizódy (najznámejší je londýnsky smog v roku 1952, ktorý si vyžiadal dokonca množstvo ľudských životov). Práve tieto problémy ukázali na dôležitosť monitorovania kvality ovzdušia. Nie je však možné umiestniť meraciu stanicu na každom mieste na svete, rovnako, ako samotné meranie nedokáže zodpovedať na otázku pôvodu znečistenia alebo procesov, ktoré ho ovplyvňujú. A preto je dôležitým doplnkom monitoringu matematické modelovanie. Matematické modely kvality ovzdušia sú (často) deterministické (spájajú príčinu s následkom), ale v mnohých praktických aplikáciách sú doplnené štatistickými metódami, ako je zhluková analýza.

Druhy modelov kvality ovzdušia sú popísané v druhej kapitole spolu krátkou charakteristikou znečisťujúcich látok a súvisiacich procesov v atmosfére. Na záver druhej kapitoly uvedieme príklady využitia zhlukovej analýzy v oblasti kvality ovzdušia.

V tretej kapitole skombinujeme teoretické základy z prvej a druhej kapitoly a aplikujeme zhlukovú analýzu na reálne dáta, ktorými sú vstupy modelu CMAQ. Pokúsime sa vytvoriť model, ktorý bude čo najlepšie zachytáva vzťah medzi vstupnými a výstupnými dátami modelu.
Analýza bude prebiehať v nasledujúcich krokoch:

1. Korelačná analýza.
2. Zhluková analýza. Porovnanie dvoch metód zhlukovania:
   (a) K-means
   (b) fuzzy C-means
3. Vyhodnotenie výsledkov zhlukovania
4. Nájdenie regresných vzťahov medzi cmaqovskými vstupmi a výstupmi pre jednotlivé zhluky.
5. Porovnanie modelov

Na záver vyhodnotíme výsledky.

1 Zhluková analýza

Zhluková analýza, tiež nazývaná segementačná analýza alebo nekontrolovaná klasifikácia, je metóda vytvárania skupín objektov, ktoré nazývame zhluky tak, aby v jednom zhluku boli objekty, ktoré sú podobné a súčasne dosť sa líšili od objektov v iných zhlukoch. Zhluková analýza je často zamieňaná s klasifikáciou dát, v ktorej sú objekty rozdelené do preddefinovaných skupín, zatiaľ čo jedným z cieľov zlukovej analýzy je nájsť vhodné skupiny do ktorých dáta rozdeliť.

Zhluková analýza je významnou častou nepriameho data-minigu. Nepriamy data-minig analyzuje a skúma veľké množstva dát, s cieľom získať užitočné informácie, ako sú vzory a závislosti (patterns) v dátach. Proces zhlukovej analýzy nie je zameraný na žiadnu konkrétnu premennú, ale hľadá vzťahy medzi všetkými premennými súčasne.

Príklad 1.1. Zhluková analýza sa v informatike používa v počítačovom videní pri segmentácii obrazu. Segmentácia obrazu je rozdelenie rôznych farieb do homogénnych skupín. Zhluková analýza sa na detekciu nepredpísaných hraníc objektov na obrázku [35] .

Príklad 1.2. Zhluková analýza sa používa aj v marketingu a ekonómii. Najširšie použitie má v prieskume trhu na rozdelenie trhu na rozumné úseky a určenie vhodných cieľových skupín pre marketingové kampane. Jedným z najvýraznejších rozdelení je rozdelenie trhu za účelom identifikovania skupín, ktoré sú náchylnejšie kupovať nové produkty. Toto rozdelenie je rôzne v závislosti od produktu a mnohých faktorov, ktoré ovplyvňujú to ako jednotlivci spravujú svoje financie, preto je to otázka zhlukovania a nie klasifikácie. [35] .

Ako vidíme z vyššie uvedených príkladov, situácie, ktorými sa zaoberáme v zhlukovej analýze, sa môžu veľmi líšiť. Vo všeobecnosti pre rôzne vstupné dáta nevieme vopred povedať, aký vzťah či štruktúra sa v dátach nachádza, a preto boli vyvinuté rôzne zhlukovacích metód, ktoré umožňujú nachádzať v dátach rôzne štruktúry. Tieto metódy nedelíme podľa použitých matematických prostriedkov, ale podľa cieľov, ku ktorým smerujú. Rozlišujeme tak metódy *hierarchického zhlukovania* a *nehierarchického zhlukovania*.

Majme množinu *X*, ktorá má n objektov. *Hierarchickým zhlukovaním* vznikajú rôzne rozklady $\Omega_s = \{C_1, C_2, ..., C_m\}$ množiny *X*, navzájom rôznych neprázdnych podmnožín množiny *X*, kde prienikom podmnožiny z rozkladu Ω_s s podmnožinou z rozkladu Ω_{s+1} (rep. Ω_{s-1}) je buď jedna z nich alebo prázdná množina, zároveň v takomto prieniku existuje aspoň jedna dvojica podmnožín.

Nehierarchickým zhlukovaním vznikne rozklad $\Omega = \{C_1, C_2, ..., C_m\}$ navzájom rôznych neprázdnych podmnožín množiny *X* tak, že prienikom každých dvoch podmnožín nie je žiadna z nich. Narozdiel od hierarchického zhlukovania, pre množiny z rôznych rozkladov, ktoré vznikajú v procese nehierarchického zhlukovania, nie je možné určiť žiaden rekurentný vzťah.

V posledných rokoch sa najmä vďaka novým možnostiam výpočtovej techniky stáva populárnejšie *zhlukovanie (segmentácia) neurónovými sieťami*, pre big dáta (veľké dátove matice) sa používa aj *zhlukovanie premenných* samostatne, alebo v kombinácii so *zhlukovaním premenných a objektov zároveň*. Na zhlukovanie premenných sa môžu použiť tie isté metódy ako pre zhlukovanie objektov, hlavným rozdielom je použitie mier podobnosti a nepodobnosti. Ako miera podobnosti pri zhlukovaní premenných sa zvyčajne používa korelácia medzi premennými. Cieľom zhlukovania premenných je vytvoriť zhluky premenných, z ktorých je možné vybrať premennú -reprezentanta, ktorá bude reprezentovať všetky premenné v danom zhluku. Zhluková analýza teda môže byť použitá na zníženie dimenzionality problému. Medzi ďalšie metódy znižovania dimenzii problému patrí metóda hlavných komponent, ktorá je popísaná v

podkapitole 1.2 a faktorová analýza ktorej sa venuje [19].

Pre zhlukovanie premenných a objektov zároveň, vznikli rôzne metódy a kombinácie metód, ako je zhlukovanie do blokov (Block clustering), ktoré vytvára K navzájom disjunktných 'blokov', alebo metódy Plaid Models a Biclustering, ktoré pre potreby analýzy biologických systémov vytvárajú zhluky, ktoré sa prekrývajú. V tejto diplomovej práci sa im viac nebudeme venovať, podrobnejšie sa im venuje napríklad [19].

1.1 Matica dát

Objekty určené na zhlukovanie sú tvorené množinou n predmetov alebo javov. Každý konkrétny objekt musí byť popísaný $p - ticou$ vopred popisaných premenných (stavov). Tieto premenné spravidla popisujeme numericky a teda aj kvalitatívnym premenným pripíšeme čísla. Každý objekt je pre zhlukovú analýzu určený $p-ticou$ stavov vybraných premenných, a tento p rozmerný vektor nazývame *identifikátorom objektu*.Pre ďalšie spracovanie dátovej matice používame ako identifikátor objektu jeho poradové číslo v množine objektov, ktorá bola vopred usporiadaná. Maticu dát teda tvoríme tak, že riadky tvorí n objektov a stĺpce p premenných. Takúto dátovú maticu je potrebné pripraviť na zhlukovú analýzu. Popíšeme ďalej premenné, ktoré môže dátová matica obsahovať, transformácie a úpravy dátovej matice, ktoré je potrebné previesť, aby bolo možné spraviť na dátovej matici zhlukovú analýzu. Viac o dátovej matici a potrebných transformáciách sa pojednáva v [2], [14] a [22] z ktorých sme čerpali pri písaní tejto pod-kapitoly.

1.1.1 Typy premenných

Premennou alebo znakom rozumieme zobrazenie množiny objektov určených na zhlukovanie do množiny premenných. Typy premenných, s ktorými sa pri zhlukovej analýze môžeme stretnúť:

- *Spojité*
- *Diskrétne*
  - nominálne - poradie nie je možné určiť (napríklad meno, pohlavie, farba očí...)
  - dichotomické (binárne, ktoré nadobúdajú len dve hodnoty) - ktoré ďalej delíme na symetrické a asymetrické.

Podľa škály merania delíme premenné:

Kvalitatívne premenné nadobúdajú slovné alebo kategorické hodnoty. Delíme ich na

- nominálne - poradie nie je možné určiť,
- ordinárne - dajú sa zoradiť.

Kvantitatívne premenné nadobúdajú číselné/numerické hodnoty. Ďalej ich delíme:

- intervalové,
- pomerové.

Nominálne premenné je pred ďalšou analýzou možné previesť na binárne. V zhlukovej anlýze budeme miesto zhluku k zavádzať binárnu premennú "patriť do zhluku k".

1.1.2 Štandardizácia dát

Štandardizácia dát je spôsob ako dáta zbaviť mierky a jednotiek. Štandardizáciou sa však stratí informácia o mierke pôvodných hodnôt dát. Pre zhlukovú analýzu, ktorá používa rôzne miery podobnosti a nepodobnosti, je štandardizácia nevyhnutným krokom. Neštandardizované dáta totiž pri použití funkcií podobnosti (nepodobnosti) alebo metriky, ako je Euklidovská metrika, pri zhlukovaní dávajú väčšiu váhu premenným, ktoré majú väčšiu mierku, čo nie je vždy

žiadúce.

Nech je daný p-dimenzionálny dataset charakterizujúci n objektov $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$. Potom matica dát je $n \times p$ daná :

$$(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)^T = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ddots & x_{np} \end{pmatrix} \tag{1.1}$$

Štandardizované hodnoty vypočítame z pôvodných dát nasledovne:

$$x^*_{ij} = \frac{x_{ij} - \overline{x}_j}{\sigma_j} \tag{1.2}$$

Kde $\overline{x}_j$ je aritmetický priemer $j-$tej premennej

$$\overline{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \tag{1.3}$$

a σ_j je smerodajná odchýlka

$$\sigma_j = \Big[\frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2\Big]^{1/2}. \tag{1.4}$$

Takto štandardizované hodnoty premenných majú strednú hodnotu 0 a rozptyl 1.

1.1.3 Korelácia

Závislosti medzi premennými môžu zhlukovú analýzu značne ovplyvniť. To, ako veľmi závislosti medzi premennými zhlukovanie ovplyvnia, závisí na tom, ako zvolený typ zhlukovania hodnotí podobnostné vzťahy objektov a zhlukov. Preto majú ideálne dáta, ktoré sú určené na zhlukovanie, navzájom nezávislé premenné. V reálnom svete sú však nezávislé dáta zriedkavé, a preto je potrebné ich vzájomnú závislosť kvantitatívne ohodnotiť.

Z hľadiska matematickej štatistiky v tomto prípade považujeme premenné za náhodné veličiny a preto s nimi môžme aj takto pracovať. Na výpočet lineárnej závislosti medzi premennými sa používa korelácia. Závislosť dvoch sledovaných premenných môže mať rôzny charakter.

Najpoužívanejší **korelačný koeficient** (Pearsonov korelačný koeficient) meria lineárnu závislosť medzi dvoma skúmanými náhodnými veličinami. Označme variáciu i-tej veličiny (premennej), ktorú počítame ako kvadrát smerodajnej odchýlky, σ_i^2, variáciu j-tej veličiny σ_j^2 a kovarianciu medzi veličinami X_i a X_j ako σ_{ij}. Kovariancia je definovaná ako $\sigma_{ij} = E(X_i - EX_i)(X_j - EX_j)$.Za predpokladu, že výber pochádza z normálneho rozdelenia, $\sigma_i^2 > 0$ a $\sigma_j^2 > 0$ je korelačný koeficient definovaný ako:

$$\rho = \frac{\sigma_{ij}}{\sqrt{\sigma_i^2 \sigma_j^2}} \tag{1.5}$$

Pre tento koeficient platí $-1 \leqslant \rho \leqslant 1$. Ak je $\rho = 0$, sú náhodné veličiny lineárne nezávislé (nekorelované). Keď je hodnota ρ blízko jednej z krajných hodnôt intervalu, tak sú náhodné veličiny lineárne závislé (korelované).

Skutočné dáta však nemusia spĺňať predpoklad normálneho rozdelenia a a závislosť medzi veličinami nemusí byť len lineárna. V takom prípade sa používa **Sprearmanov korelačný koeficient**, ktorý sa definuje ako výberový korelačný koeficient poradia hodnôt objektov v danej premennej. Nech $R_1, ..., R_n$ je vektor poradia hodnôt pre vzostupne zoradené hodnoty prvej premennej a $Q_1, ..., Q_n$ je vektor poradia hodnôt pre druhú premennú. Takýto koeficient potom neskúma len lineárnu závislosť medzi premennými ale aj ľubovoľnú inú monotónnu závislosť. Spearmanov korelačný koeficient sa dá prepísať aj ako viz.[2].

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2 \tag{1.6}$$

1.1.4 Analýza hlavných komponentov

Analýza hlavných komponentov (anglicky principal component analysis PCA) je používaná najmä na redukciu dimenzii pri veľko dimenzionálnych dátach, pričom zachováva čo najviac variability originálnych dát. Nové premenné, ktoré vzniknú transformáciou dát sa nazývajú hlavné komponenty. Sú *nekorelované* a zoradené tak, že prvá komponenta zahŕňa najviac rozptylu z pôvodných dát a posledná najmenej. Hlavné komponenty sú definované nasledovne:

Nech $\mathbf{v} = (v_1, v_2, ..., v_p)'$ je vektor p náhodných premenných. Prvý krok je nájsť lineárnu funkciu $\mathbf{a}_1'\mathbf{v}$ prvkov $\mathbf{v}$, ktorá maximalizuje rozptyl, kde $\mathbf{a}_1$ je p dimenzionálny vektor $(a_{11}, a_{12}, ..., a_{1p})$:

$$\mathbf{a}_1'\mathbf{v} = \textstyle\sum_{i=1}^{d} a_1 v_i.$$

Po nájdení funkcií $\mathbf{a}_1'\mathbf{v}$, $\mathbf{a}_2'\mathbf{v}$, ..., $\mathbf{a}_{j-1}'\mathbf{v}$, hľadáme ďalšiu funkciu $\mathbf{a}_j'\mathbf{v}$, ktorá je ortogonálna s funkciami $\mathbf{a}_1'\mathbf{v}$, $\mathbf{a}_2'\mathbf{v}$, ..., $\mathbf{a}_{j-1}'\mathbf{v}$ a má maximálny možný rozptyl. Teda po d krokoch nájdeme d lineárnych funkcií. Takto odvodená j-tá premennú $\mathbf{a}_j'\mathbf{v}$ je j-tý hlavný komponent (PC).

Na nájdenie vhodných lineárnych funkcií je potrebná kovariačná matica Σ vektoru v. Pre väčšinu reálnych prípadov je kovariačná matica neznáma a je nahradená odhadom kovariačnej matice. Pre $j = 1, 2, ..., p$ sa dá ukázať že j-ty PC je daný ako $z_j = \mathbf{a}_j'\mathbf{v}$ kde $\mathbf{a}_j$ je vlastný vektor kovariačnej matice Σ zodpovedajúci $j - tej$ najväčšej hodnote vlastného čísla λ_j.

Naozaj, v prvom kroku môžeme $z_1 = \mathbf{a}_1'\mathbf{v}$ nájsť riešením nasledujúcej optimalizačnej úlohy :

$$\max \sigma(\mathbf{a}_1'\mathbf{v}) \text{ subject to } \mathbf{a}_1'\mathbf{a} = 1,$$

kde $\sigma(\mathbf{a}_1'\mathbf{v})$ je počítaný ako :

$$\sigma(\mathbf{a}_1'\mathbf{v}) = (\mathbf{a}_1'\Sigma\mathbf{a}_1).$$

Na riešenie tohto optimalizačného problému sa používa metóda Lagrangeových multiplikátorov. Nech λ je Lagrangeov multiplikátor. Budeme maximalizovať :

$$\mathbf{a}_1'\Sigma\mathbf{a}_1 - \lambda(\mathbf{a}_1'\mathbf{a} - 1). \tag{1.7}$$

Derivovaním rovnice (1.7) podľa $\mathbf{a}_1$ dostávame :

$$\Sigma\mathbf{a}_1 - \lambda\mathbf{a}_1 = 0,$$

alebo

$$(\Sigma - \lambda I_d)\mathbf{a}_1 = 0,$$

kde I_d je identická matica $d \times d$.

Teda λ je vlastné číslo matice Σ a $\mathbf{a}_1$ je k nemu zodpovedajúci vlastný vektor. Keďže

$$\mathbf{a}_1'\Sigma\mathbf{a}_1 = \mathbf{a}_1'\lambda\mathbf{a}_1 = \lambda,$$

$\mathbf{a}_1$ je vlastný vektor prislúchajúci k najväčšiemu vlastnému číslu Σ.

1.2 Miery podobnosti a nepodobnosti

Pre zklukovú analýzu sú dôležitými pojmami podobnosť objektov a miera podobnosti medzi objektami. Preto uvedieme nasledujúce definície z [22].

Definícia 1.3. *Funkcia podobnosti*
Funkcia $s : R^d \times R^d \to R$ sa nazýva funkcia podobnosti, ak spĺňa nasledujúce vlastnosti:
1.Symetrickosť $s(\mathbf{x}_i, \mathbf{x}_j) = s(\mathbf{x}_j, \mathbf{x}_i)$
2.Ohraničenosť zdola. Pre väčšinu funkcii podobnosti to znamená nezápornosť

$$\forall \mathbf{x}_i, \mathbf{x}_j : 0 \leq s(\mathbf{x}_i, \mathbf{x}_j)$$

, ale pre niektoré, ako je napríklad korelácia je podmienka tvaru

$$\forall \mathbf{x}_i, \mathbf{x}_j : -1 \leq s(\mathbf{x}_i, \mathbf{x}_j)$$

3.Pre $(\mathbf{x}_i = \mathbf{x}_j)$ prestavuje hodnota $s(\mathbf{x}_i, \mathbf{x}_j)$ maximálnu hodnotu z odboru hodnôt s.

Podmienky funkcie podobnosti spĺňa aj korelačný koeficient.

Definícia 1.4. *Funkcia nepodobnosti*
Funkcia $d : R^d \times R^d \to R$ sa nazýva funkcia nepodobnosti, ak spĺňa nasledujúce vlastnosti:
1.$d(\mathbf{x}_i, \mathbf{x}_j) = 0 \iff (\mathbf{x}_i = \mathbf{x}_j)$,
2.Nezápornosť $\forall \mathbf{x}_i, \mathbf{x}_j : 0 \leq d(\mathbf{x}_i, \mathbf{x}_j)$
3.Symetrickosť: $d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i)$.

Definícia 1.5. *Metrika*
Na množine X definujeme funkciu $d : X^d \times X^d \to R$, ktorá spĺňa nasledujúce vlastnosti:
1. $d(\mathbf{x}_i, \mathbf{x}_j) = 0 \iff (\mathbf{x}_i = \mathbf{x}_j)$,
2. Nezápornosť: $\forall \mathbf{x}_i, \mathbf{x}_j : d(\mathbf{x}_i, \mathbf{x}_j) \geq 0$.
3. Symetrickosť: $d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i)$
4. Trojuholníková nerovnosť:$\forall \mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k : d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_k) + d(\mathbf{x}_j, \mathbf{x}_k)$,
sa nazýva metrika.

Medzi najznamejšie typy metriky patrí euklidovska d_E,vážená euklidovská d_{EW}, kvadrát euklidovskej vzdialenosti d_{ES},Čebyševova d_C, Minkowského d_M, Manhattanská d_B

$$d_E(\mathbf{x_i}, \mathbf{x_j}) = \sqrt{\sum_{l=1}^{p}(x_{il} - x_{jl})^2} = \|\mathbf{x_i} - \mathbf{x_j}\| \tag{1.8}$$

$$d_{EW}(\mathbf{x_i}, \mathbf{x_j}) = \sqrt{\sum_{l=1}^{p} w_l^2(x_{il} - x_{jl})^2} \tag{1.9}$$

$$d_{ES}(\mathbf{x_i}, \mathbf{x_j}) = \sum_{l=1}^{p} (x_{il} - x_{jl})^2 \tag{1.10}$$

$$d_C(\mathbf{x_i}, \mathbf{x_j}) = max_l(|x_{il} - x_{jl}|) \tag{1.11}$$

$$d_B(\mathbf{x_i}, \mathbf{x_j}) = \sqrt[q]{\sum_{l=1}^{p} |x_{il} - x_{jl}|^q} \tag{1.12}$$

$$D_C(\mathbf{x_i}, \mathbf{x_j}) = \sum_{l=1}^{p} |x_{il} - x_{jl}| \tag{1.13}$$

1.3 Zhluk

Jedným zo základných pojmov zhlukovej analýzy je zhluk. Existuje mnoho definícii zhluku a žiadna z nich nie je ustálená. Rôzne metódy používajú rôzne definície zhluku, čo má za následok že výsledné zhluky medzi metódami sa môžu líšiť. V tejto diplomovej práci ako zhluk rozumieme definíciu podla [22]:

Definícia 1.6. Je daná množina objektov $\mathbf{O} = \{O_1, O_2, ..., O_n\}$ a funkcia nepodobnosti objektov d. *Zhlukom* nazveme takú podmnožinu *C* množiny objektov **O**, pre ktorú platí :

$$\max_{O_i, O_j \in C} d(O_i, O_j) < \min_{\substack{O_i \in C; \\ O_k \notin C}} d(O_i, O_k) \tag{1.14}$$

Pre takto definované zhluky je následne potrebné zaviesť koeficient nepodobnosti zhlukov ktorý umožní kvantitatívne ohodnotenie podobnostných vzťahov zhlukovania.Tieto vzťahy sme čerpali z [22]. V nasledujúcich vzťahoch budu $C_i = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_r\}$ a $C_j = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_s\}$ označovať dva zhluky veľkosti *r* a *s* z rozkladu v uvedenom poradí.

Definícia 1.7. Funkcionál **D**, ktorý priraďuje každej dvojici (C_i, C_j) v rozklade $\Omega = \{C_1, C_2, ..., C_m\}$ číslo $\mathbf{D}(C_i, C_j)$ sa nazýva *koeficient nepodobnosti zhlukov rozkladu* ak spĺňa nasledujúce podmienky :

$$\begin{aligned} &\mathbf{D}(C_i, C_i) = 0 \\ &\mathbf{D}(C_i, C_j) \geqslant 0 \\ &\mathbf{D}(C_i, C_j) = \mathbf{D}(C_j, C_i) \end{aligned} \tag{1.15}$$

Ďalej uvedieme najpoužívanejšie funkcionály (metódy), ktoré spĺňajú vyššie uvedené podmienky.

Metóda najbližšieho suseda Nech $d(\cdot, \cdot)$ je funkcia nepodobnosti, potom táto metóda definuje vzdialenosť zhlukov ako medzi C_i a C_j ako :

$$D_{nn}(C_i, C_j) = \min_{1 \leqslant k \leqslant r, 1 \leqslant l \leqslant s} d(\mathbf{y}_k, \mathbf{z}_l) \tag{1.16}$$

Metóda najvzdialenejšieho suseda Nech $d(\cdot, \cdot)$ je funkcia nepodobnosti, potom táto metóda definuje vzdialenosť zhlukov ako medzi C_i a C_j ako :

$$\begin{aligned} &D_{fn}(C_i, C_j) = \max_{1 \leqslant k \leqslant r, 1 \leqslant l \leqslant s} d(\mathbf{y}_k, \mathbf{z}_l) \\ &D_{fn}(C_i, C_i) = 0 \end{aligned} \tag{1.17}$$

Centroidna metóda Jednou z najpoužívanejších metód na meranie nepodobnosti zhlukov je centroidná metóda tiež známa ako metóda vzdialenosti priemerov, ktorá vyjadruje nepodobnosť dvoch zhlukov ako vzdialenosť ich centroidov, ktoré tiež nazývame ťažiská. Pre dva zhluky C_i a C_j je táto metóda definovaná ako:

$$D_c(C_i, C_j) = d(\overline{C_i}, \overline{C_j}) \tag{1.18}$$

kde priemery zhlukov - ťažiská sú počítane nasledovne:

$$\begin{aligned} \overline{C_i} &= \frac{1}{r} \sum_{\mathbf{y} \in C_i} \mathbf{y} \\ \overline{C_j} &= \frac{1}{s} \sum_{\mathbf{z} \in C_j} \mathbf{z} \end{aligned} \tag{1.19}$$

1.4 Hierarchické zhlukovanie

Metódy hierarchického zhlukovania delíme na *aglomeratívne* a *divízne*. Aglomeratívne zhlukovacie metódy, tiež nazývané "zdola-hore"metódy, začínajú tak, že každý objekt sa nachádza v samostatnom zhluku, potom postupne na základe vopred definovaného pravidla zhluky spájame, až kým nemáme jeden zhluk, ktorý obsahuje všetky objekty. Divízne zhlukovacie metódy, prezývané "zhora-dole"metódy, fungujú presne naopak. Objekty sú najprv zaradené v jedom zhluku, ktorý sa potom rozdelí na dva zhluky, a tak ďalej až kým každý objekt nie je v samostatnom zhluku. Literatúra sa viac venuje aglomeratívnym zhlukovacím metódam, objavili sa však mnohé argumenty, ktoré tvrdia že divízne metódy poskytujú sofistikovanejšie a robustnejšie zhluky. Hierarchickým metódam zhlukovania sa detailnejšie venujú [19], [22]

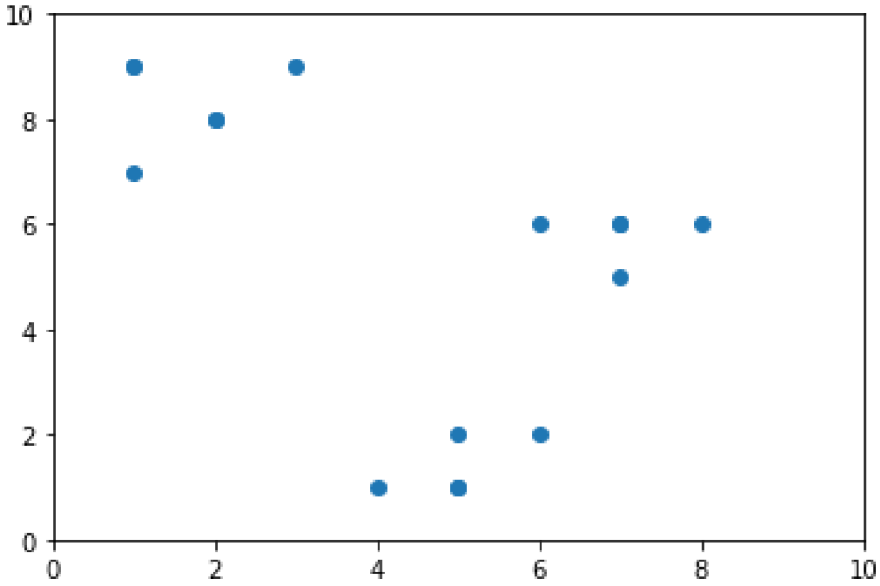
1.4.1 Dendogram

Konečným výsledkom všetkých hierarchických zhlukovacích metód je dendogram (teda hierarchicky stromový diagram), v ktorom riešenie pre k zhlukov získame spojením nejakých zhlukov z riešenia pre $(k + 1)$ zhlukov.
Vertikálny dendogram zobrazuje 'výšku' spájacieho kritéria, na ktorej sú objekty alebo zhluky alebo oboje spojené a vytvárajú nový väčší zhluk. Objekty, ktoré sú si podobné, sú spájané v malej výške, zatiaľ čo objekty, ktoré sú si viac menej nepodobné sú spojené v dendograme až vyššie. Podobnosť objektov v dendograme teda čítame z výšky ich spojenia.

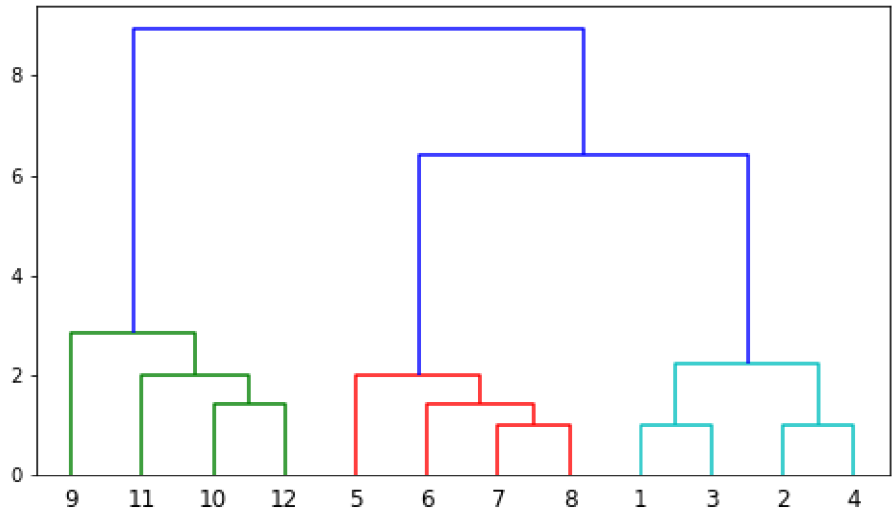
Rozdelenie dát do špecifického počtu zhlukov je teda možné 'odstrihnutím' dendogramu vo vhodnej výške. Keď v dendograme v nejakej výške zaznačíme horizontálnu čiaru, potom číslo K, vertikálnych čiar ktoré boli preťaté touto čiarou, identifikuje K zhlukové riešenie. Každé preťatie identifikuje jeden zhluk a objekty, ktoré sú na konci tejto vetvy, sú objektami daného zhluku.

Výpočtový software, ktorý použijeme na vytvorenie dendogramu je vo všeobecnosti napísaný tak, aby bol dendogram ľahko interpretovateľný. Pre veľké dátové matice sa však tento cieľ stáva nemožný.

Príklad 1.8. Dendogram
Majme objekty dané nasledujúcim zápisom, (x,y)=[(5;2),(4;1),(5;1),(6;2),(8;6),(6;6),(7;6),(7;5),(1;7),(2;8),(1;9),(3;9)], zobrazené na obr.1.
Pre tieto objekty vznikne použitím najvzdialenejšieho suseda dendogram na obr.2.



Obr. 1: Zobrazenie dátových bodov



Obr. 2: Dendogram použitím metódy najvzdialenejšieho suseda

1.4.2 Aglomeratívne zhlukovacie metódy

Medzi najpopulárnejšie aglomeratívne zhlukovacie metódy patria metóda najbližšieho suseda (jednoduché spojenie), metóda najvzdialenejšieho suseda (kompletné prepojenie) a centroidná metóda (priemerné spojenie). Tieto metódy definujú spôsob akým sú dva zhluky (ktoré môžu obsahovať len jediný prvok) spojené a vytvoria väčší zhluk.

Žiaden z týchto algoritmov nie je univerzálne najlepší pre všetky situácie. Metóda najbližšieho suseda často vedie k dlhým reťaziam zhlukov, ktoré sú spojené samostatnými objektami, ktoré sú blízko seba, takýto výsledok je v praxi mnohokrát nežiaduci. Metóda najvzdialenejšieho suseda, tvorí viacero menších veľmi kompaktných zhlukov. Centroidná metóda na rozdiel od metódy najbližšieho suseda a metódy najvzdialenejšieho suseda, závisí na veľkosti zhluku.

1.4.3 Divízne zhlukovacie metódy

Spomeňme len najpoužívanejšiu metódu divízneho hierarchického zhlukovania. Táto metóda spočíva v tom že v každom kroku objekty rozdelíme na "splinter"(úlomok), ktorý nazveme zhluk A, a zvyšok, ktorý nazveme zhluk B. Objekt, ktorý má najväčšiu priemernú nepodobnosť s ostatným objektami v množine zhlukovaných predmetov, zvolíme na úvod ako splinter, tento objekt vytvorí zhluk A. Po tomto úvodnom rozdelení dát do zhlukov A a B, vypočítame pre každý objekt v zhluku B nasledujúce hodnoty:

1. priemernú nepodobnosť medzi objektom a všetkými ostatnými objektami v zhluku B
2. priemernú nepodobnosť medzi objektom a všetkými objektami v zhluku A.

Následne vypočítame rozdiel 1.-2. pre každý objekt v zhluku B. Ak sú všetky tieto rozdiely záporné, algoritmus zastavíme. Ak sú niektoré z týchto rozdielov pozitívne, znamená to, že daný objekt v zhluku B je v priemere podobnejší objektom v zhluku A, ako ostatným objektom zhluku B. Objekt s najväčším pozitívnym rozdielom následne premiestnime zo zhluku B do zhluku A, a proces opakujeme. Týmto algoritmom vznikne binárne rozdelenie na dva zhluky. Taký istý proces môžeme následne aplikovať na vzniknuté zhluky a vytvoriť tak aj z nich binárne rozdelenie.

1.5 Nehierarchické zhlukovanie

Nehierarchickým zhlukovacím metódam, tiež nazývaným *metódy rozkladu*, sa táto diplomová práca bude venovať viac, keďže práve tieto metódy poskytujú možnosť využiť aj fuzzy množiny. Tieto metódy rozdelia dáta do vopred určeného počtu zhlukov K. Medzi riešením ktoré poskytuje K zhlukov a riešením pre $K + 1$ zhlukov, neexistuje žiaden hierarchický vzťah. Pre dané K hľadáme také rozdelenie do K samostatných zhlukov, ako sú definované v def.1.6.

Jedna z možných metód by mohla byť taká, že by sme v prvom kroku vypočítali všetky možné skupiny K zhlukov objektov. Potom by sme na základe nejakého optimalizačného kritéria vybrali "najlepšiu"skupinu"zhlukov, ktorá optimalizuje dané kritérium. Pre veľké množiny dát sa očividne takýto prístup veľmi rýchlo stáva neriešiteľným a vyžadoval by obrovské množstvo energie, počítačového času a úložného priestoru. Kvôli tomu sú všetky používané zhlukovacie algoritmy iteratíve a počítajú sa len s malým množstvom vyčíslení rôznych funkcii podobností alebo nepodobnosti.[19]

1.5.1 K-means

Jedna z najpopulárnejších nehrierachických metód sa nazýva K-means, teda metóda K-priemerov. Je populárna najmä vďaka jednoduchosti jej algoritmu. Jej popis je zhrnutý v nasledovných krokoch z [35]:

1. Na úvod sa zvolí K počiatočných ťažísk $m_i, ..., m_K$ tvoriacich maticu M buď celkom náhodne, alebo podľa určitej predošlej znalosti. Táto počiatočná matica je jedna zo slabých stránok k-means algoritmu. Je od nej závislé celkový výsledok algoritmu a ak je správne zvolená, algoritmus môže skončiť v lokálnom optime.

2. Metódou najbližšieho suseda priradíme objekty k jednotlivým centroidom a vytvoríme tak k zhlukov. Objekt x_j je zaradený týmto algoritmom do zhluku so stredom m_i práve vtedy, keď $\forall l \in 1, ..., K, i \neq l : \|x_j - m_i\| < \|x_j - m_l\|$.

3. Podľa aktuálneho zadelenia objektov prerátame centroidy zhlukov podľa:

$$m_i = \frac{1}{n_i} \sum_{x_j \in C_i} x_j \tag{1.20}$$

čím vznikne nová matica M

4. Porovnáme novú maticu M s predošlou maticou M. Ak je nová matica iná, je potrebné zopakovať kroky 2 a 3 znova, aby sa centroidná matica ustálila a tým pádom aj zadelenie objektov ku centroidom, ktoré reprezentujú jednotlivé zhluky, by bolo definitívne. V prípade, že sa matica M nezmenila, ukončíme algoritmus a vyhodnotíme výsledky.

Bod 4 sa v prípade väčšieho množstva dát dá spraviť nasledovne

- Spočítame rozdiel medzi novou maticou M a pôvodnou maticou M. Ak je tento rozdiel väčší ako ľubovoľne zvolené kladné ϵ, je potrebné zopakovať kroky 2 a 3 znova, aby sa centroidná matica ustálila a aby aj zadelenie objektov k ťažiskom, ktoré reprezentujú jednotlivé zhluky bolo definitívne. V prípade, že je rozdiel medzi maticami menší ako ϵ, ukončíme algoritmus a vyhodnotíme výsledky.

Optimálny počet zhlukov

Jednou z dôležitých informácii je zvolenie počtu zhlukov. V mnohých reálnych prípadoch máme nejakú vedomosť o dátach, ktorá nám umožní zvolenie vhodného počtu zhlukov.Ale nemusí to tak byť, v takom prípade je potrebné nájsť dátovú štruktúru priamo z dátovej matice a nájsť optimálny počet zhlukov. Existuje množstvo metód hodnotiacich optimálny počet zhlukov pre k-means metódu. Z heurustického hľadiska najvhodnejšou a dlho najpoužívanejšou metódou je 'elbow method'=lakťová metóda. Táto metóda používa funkcionál sumy súčtu štvorcov odchýliek objektov zhluku(n_j je počet objektov patriacich do zhluku) od ťažiska zhluku rozkladu pre všetky zhluky

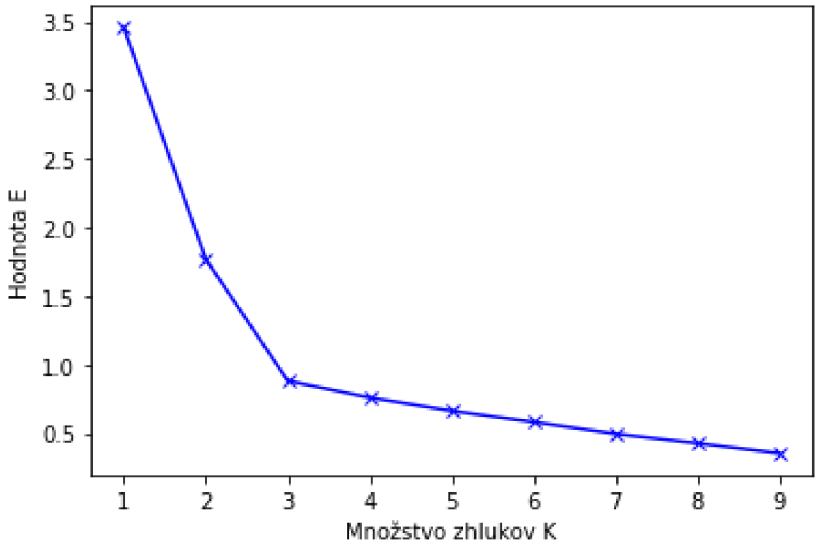
$$E = \sum_{j=1}^{K} \sum_{i=1}^{n_j} d_{ES}(x_{i,j}, m_j). \tag{1.21}$$

Metóda spočíva v zhotovení k-means rozkladu a zistení hodnôt tohto funkcionálu pre hodnoty množstva zhlukov K od 1 po vopred zvolené množstvo zhlukov l, kde l je maximálny počet zhlukov, ktorý je rozumné uvažovať pre dané dáta. Tieto hodnoty pre rôzne počty zhlukov sú následne nanesené do grafu, na ktorom sú počty zhlukov a k ním priradené hodnoty funkcionálov, ktorými preložíme krivku. V takto zhotovenom grafe následne hľadáme 'lakeť', v ktorom sa prudko táto krivka zmení a túto hodnotu k zvolíme ako počet zhlukov. Hlavnou myšlienkou za týmto výberom je klesanie hodnoty E pre zvyšujúci sa počet zhlukov K, hodnota E bude klesať až pre $K = n$ bude platiť $E = 0$. Ak je v dátach nejaká štruktúra zhlukov, ktorú je možné zhlukovaním odhaliť, tak bude hodnota funkcionálov pre rôzne K rýchlo klesať do počtu zhlukov, ktoré sa v tejto štruktúre nachádzajú, následne po prekročení tejto hodnoty klesanie nebude tak prudké, lebo dochádza k deleniu existujúcich zhlukov, a teda odchýlka od ťažiska sa už neklesá tak významne. Takže cieľom je určiť pomerne nízky počet zhlukov ktorý má čo najnižšiu hodnotu funkcionálu E, a to je práve nájdením 'lakťa'. Čerpali sme zo zdrojov [16] a [22].

Príklad 1.9. Lakťová metóda

Majme body (x,y)=[(5;2),(4;1),(5;1),(6;2),(8;6),(6;6),(7;6),(7;5),(1;7),(2;8),(1;9),(3;9)], ako pre

príklad 1.8
Zobrazenie závislosti funkcionálu E na počte zhlukov vyzerá nasledovne
Krivka jednoznačne mení spád pre $K = 3$ a práve tu môžme vidieť aj 'lakeť'. Aj z dát môžeme



Obr. 3: Lakťová metóda

pozorovať, že sa v nich nachádzajú 3 zhluky.

Pre veľké množstvo dát však nemusí byť zobrazenie objektov do bodového grafu, ako je na Obr.1, a rozlíšenie zhlukov používateľom možné. Hodnotu funkcionálu E však bude v takom prípade možné vypočítať výpočtovým softwarom a bude možné použiť lakťovú metódu.

1.5.2 Fuzzy c-means

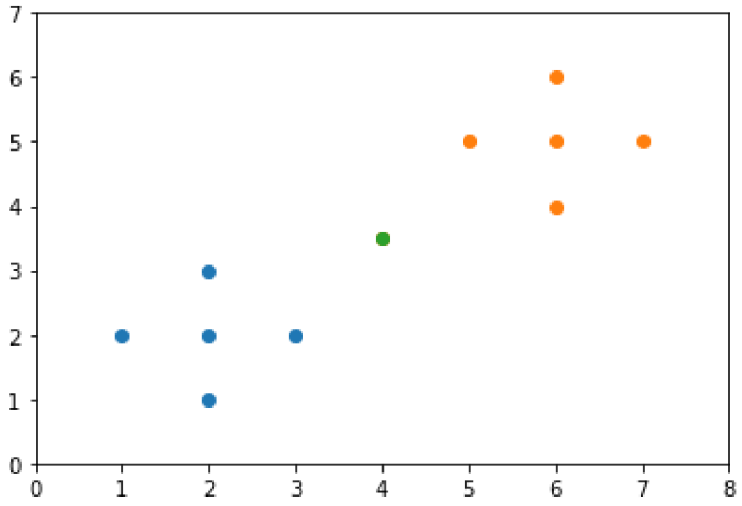
Mnohé dáta však nie je možné jednoznačne rozdeliť do zhlukov. Na Obr.4 je zobrazená presne takáto situácia. Zatiaľ čo o modrých a oranžových bodoch sa dá jednoznačne povedať do akých zhlukov patria, pri zelenom to nie je jasné. Pri tradičnom delení pomocou metódy k-means by bol pri delení na dva zhluky pridelený do zhluku na základe počiatočnej polohy centroidov. Zároveň by delenie nebolo optimálne, keďže hodnota funkcionálu 1.21 by bola pri rozdelení na dva zhluky značne vyššia, ako pri rozdelení na tri zhluky, vytvorenie tretieho zhluku však nemusí byť vždy žiaduce.

Túto situáciu rieši fuzzy c-means zhlukovanie, ktoré umožňuje bodom patriť do viacerých zhlukov. V príklade z obr.4 umožňuje zelenému bodu patriť modrému aj oranžovému zhluku. Táto metóda je založená na fuzzy logike, ktorej základom je definícia fuzzy definície podla Zadeha [37]. Okrem Zadeha sa téme fuzzy množín venujú [21], [26], [24], z ktorý sme čerpali.

Definícia 1.10. Fuzzy množina:
Nech $X \neq \emptyset$ je pevne zvolená univerzálna množina. *Fuzzy podmnožinou* A univerzálnej množiny X (stručne *fuzzy množinou*), budeme rozumieť objekt popísaný (zovšeobecnenou) charakteristickou funkciou

$$\mu_A : X \to \langle 0, 1 \rangle,$$

ktorá sa tiež nazýva *funkcia príslušnosti*. Pre každý prvok $x \in X$ hodnota $\mu : X \to \langle 0, 1 \rangle$, je stupeň príslušnosti prvku x do fuzzy množiny A. Stupeň príslušnosti je teda stotožnení s fuzzy množinou a naopak.



Obr. 4: Nejasne rozdelenie zhlukov

Pre takto definované množiny je potrebné definovať logické spojky, ktoré sú založené na funkcii príslušnosti. Takéto logické spojky môžu byť definované rôznymi spôsobmi. V tejto práci spomenieme Lukasiewiczovu logiku, v ktorej sa pravdivostné hodnoty zložených tvrdení, pre fuzzy množiny A a B počítajú pomocou nasledujúcich vzťahov

$$\mu_{(A\wedge B)} = min(\mu_A, \mu_B),$$

$$\mu_{(A\vee B)} = max(\mu_A, \mu_B),$$

$$\mu_{(\neg A)} = 1 - \mu_A,$$

$$\mu_{(A\Rightarrow B)} = min(1, 1 - \mu_A + \mu_B),$$

$$\mu_{(A\Leftrightarrow B)} = 1 - |\mu_A + \mu_B|.$$

Algoritmus fuzzy c-means

Fuzzy c-means (FCM) je algoritmus, umožňuje prvkom patriť do viacerých zhlukov s istým stupňom príslušnosti. To je veľmi užitočné, keď majú zhluky nejasné hranice ako na Obr.4, alebo nie sú zhluky jednoznačne rozdelené. Okrem toho môže stupeň príslušnosti pomôcť užívateľom objaviť zložitejšie vzťahy medzi jednotlivými objektami a zhlukmi. Použitiu fuzzy množín sa podrobnejšie okrem literatúry spomenutej vyššie venujú šlánky [3], [35] a [7], z ktorých sme tiež čerpali pri tvorbe tejto pod-kapitoly. Metóda spočíva v minimalizácii funkcie.

$$J_m = \sum_{i=1}^{n}\sum_{j=1}^{p} u_{ij}^m d_{ES}(x_i, c_j), \tag{1.22}$$

Kde c je počet zhlukov,

- m je váhový exponent $m \in \langle 1, \infty)$. Pre väčšie hodnoty m sú zhluky viac fuzzy zatiaľ čo pre $m = 1$ sa jedná o k-means zhlukovanie. Váhový exponent je jeden z najdôležitejších parametrov FCM, aj napriek tomu neexistuje žiaden návod ako ho nastaviť. V mnohých praktických situáciach sa odporúča zvoliť m rovné 2. Bezdek [28] odporúča odporúča interval $\langle 1,5; 2\rangle$. V niektorých zdrojoch ako spodnú hranicu uvádzajú $m \geq \frac{n}{(n-2)}$ [4] a túto hodnotu aj Bezdek a Pal [28] označili za prvý teoretický výsledok, ktorý

usmerňuje hľadanie optimálneho váhového exponentu. Ďalším teoretickým výsledkom, ktorý umožňuje presnejšie určenie váhového exponentu je $m > (1/(1 - 2\lambda_{max}(F_{U^*})))$ ak $\lambda_{max}(F_{U^*}) < 0,5$, ktorému sa ďalej venuje článok [36]

- u_{ij} je stupeň príslušnosti pre každý objekt x_i do j-tého zhluku, pre ktorý platia nasledujúce podmienky:

$$u_{ij} \in \langle 0, 1 \rangle, i = 1, ..., n; j = 1, ..., p$$

$$\sum_{j=1}^{p} u_{ij} = 1, i = 1, ..., n, \tag{1.23}$$

Podmienka 1.23 zaručuje že všetky objekty majú v zhlukovaní dohromady rovnakú váhu.

$$0 < \sum_{i=1}^{n} u_{ij} < N, j = 1, ..., p \tag{1.24}$$

Podmienka 1.24 zaručuje, že žiaden zhluk nie je prázdny.

- x_i i-tý m-rozmerný prvok dátovej matice

- c_j p-rozmerné ťažisko zhluku.

Základné kroky FCM algoritmu sa zhrnieme nasledovne:

1. Zvolíme vhodné hodnoty ako počet zhlukov c a váhový exponent m, a ľubovoľne malé a kladné ϵ. Náhodne generujeme vstupnú maticu **C**, táto matica ma rovnakú úlohu ak v k-means algoritme. Tiež ovplyvňuje výsledné zhluky a môže spôsobiť, že sa algoritmus skončí v lokálnom minime. A premennú t položíme $t = 0$.
2. Výpočet novej matice **U** ako

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} (\frac{\|\mathbf{x}_i - c_j\|}{\|\mathbf{x}_i - c_k\|})^{\frac{2}{d-1}}}, i = 1, 2, ...n, j = 1, 2, ...c; \tag{1.25}$$

3. Výpočet novej matice **C** ako

$$c_j^{t+1} = \frac{\sum_{i=1}^{n} u_{ij}^m \mathbf{x}_i}{\sum_{i=1}^{n} u_{ij}^m} \tag{1.26}$$

4. Kroky 2 a 3 opakujeme až kým nie je splnená podmienka $\|C^{t+1} - C^t\| < \epsilon$

Fuzzy Partition Coefficient

Na vyhodnotenie vlastností fuzzy zhlukovania existuje množstvo rôznych koeficientov, ktorých prehľad je možné nájsť [35]. Najpoužívanejší je *fuzzy partiton coeficient* (fpc) - koeficient fuzzy rozdelenia, ktorý bol navrhnutý Bezdekom. Na jeho výpočet sa používa len matica príslušností **U** a je definovaný nasledovne:

$$fpc = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{K} u_{ij}^2. \tag{1.27}$$

Takto definovaný koeficient fpc nadobúda hodnoty z intervalu $[1/K, 1]$. Ak je koeficient rovný 1, tak zhlukovanie zodpovedá pevným zhlukom, ak je koeficient rovný $1/K$, tak objekty patria všetkým zhlukom rovnako. V takom prípade to znamená, že zhlukovací algoritmus nie je schopný nájsť štruktúru zhlukov, alebo že v dátach žiadna takáto štruktúra nie je.

Hodnota fpc je ovplyvnená váhovým exponentom m. Keď je m približne rovné 1, tak aj koeficient fpc je približne 1, nezávislo na množstve zhlukov K.

1.6 Zhluková analýza použitím neurónových sieti

Neurónové siete sa v poslednej dobe tešia stúpajúcej popularite na riešenie rôznych úloh a problémov, najme vďaka vývoju počítačového hardware, počítačových systémov a zvýšeniu výpočtovej kapacity. Na vytvorenie zhlukov sa používajú neurónové siete, ktoré sú založené na súťaživom učení (competitive learning). O zhlukovej analýze použitím neurónových sieti sa venujú kapitoly v [35] a [11] z ktorých sme čerpali aj pri tvorbe tejto pod-kapitoly.

Zmeňme formuláciu základnej úlohy zhlukovej analýzy, abe sme mohli popísať súťaživé učenie a zhlukovú analýzu popíšme ako vektorovú kvantifikáciu. Vektorová kvantifikácia je klasická metóda, ktorá sa používa na odhadovanie funkcie hustoty spojitej pravdepodobnosti náhodnej premennej $p(\mathbf{x})$ premennej $\mathbf{x} \in \mathbf{R}^p$ použitím konečného množstva prototypov (ťažísk zhlukov). Množina vektorov $\mathbf{x}$ je reprezentovaná konečnou množinou prototypov $c_1, ..., c_K \subset \mathbf{R}^p$ (ťažiskom zhluku), ktorá sa zvykne nazývať codebook. Práve codebook, je možné získať zhlukovaním. Keď je codebook špecifikovaný, odhadom $\mathbf{x}$ je vektor $\mathbf{c}_i$, ktorý je najbližšie k $\mathbf{x}$. Použitím metódy najbližšieho suseda pri hľadaní codebooku sa tento proces sa nazýva jednoduché súťaživé učenie (SCL).

Codebook môže byť nájdený minimalizovaním štvorcu strednej hodnoty vyčíslenia chyby

$$E = \int |\mathbf{x} - \mathbf{c}|^2 p(\mathbf{x}) d\mathbf{x} \tag{1.28}$$

kde $\mathbf{c}$ je funkcia $\mathbf{x}$ a $\mathbf{c}_j$. Pre objekt $\mathbf{x}_t$ môžme vytvoriť iteratívny vzorec na výpočet codebooku, ktorý bol odvodený Kohenonom v roku 1997

$$\mathbf{c}_j(t+1) = \mathbf{c}_j(t) + \eta(t)\delta_{\omega j}[\mathbf{x}_t - \mathbf{c}_j(t)]$$

kde index w odpovedá vyhrávajúcemu prototypu, ktorý je najbližšie k $\mathbf{x}_t$,

$\delta_{\omega i}$ je Kronekerové delta, $\delta_{\omega i}$ ma hodnotu 1 keď $\omega = i$ a inak 0

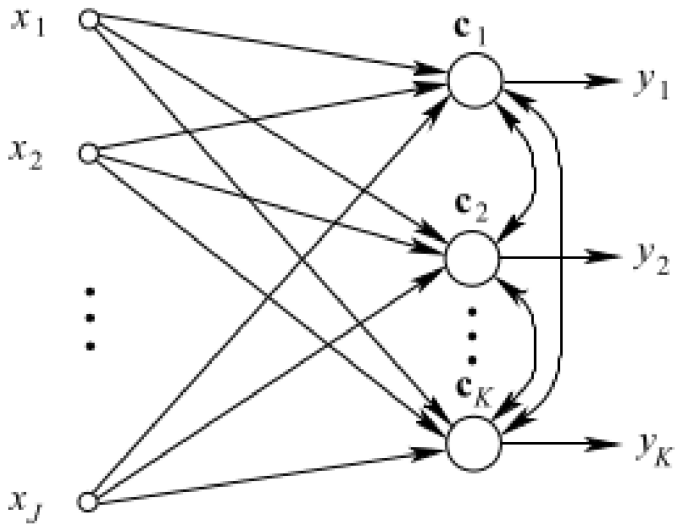
$\eta > 0$ je koeficient učenia s nízkou hodnotou, ktorý spĺňa klasické Robbins-Monro podmienky, teda $\sum \eta(t) = \infty$ a $\sum \eta(t)^2 < \infty$. Zvyčajne, η je určený tak že v čase monotónne klesá. Zvyčajne je η

Súťaživé učenie je implementované použitím dvojvrstvovej ($J - K$) neurónovej siete ako je zobrazená na Obr.5. Vstupná a výstupná vrstva sú úplne prepojené. Výstupná vrstva sa nazýva vrstva súťaženia, v ktorej sa bočné prepojenia používajú na bočnú inhibíciu.

Pre zhlukovú analýzu je štruktúra súťaživého učenia zvyčajne odovdená minimalizáciou funkcionálu súčtu štvorcov rezíduí (RSS)

$$E = \frac{1}{n} \sum_{l=1}^{n} E_l \tag{1.29}$$

$$E_l = \sum_{k=1}^{K} \mu_{kl} \|\mathbf{x}_l - c_k\|^2, \tag{1.30}$$



Obr. 5: Architektúra neurónovej siete súťaživého učenia. Prevzaté z [11].

kde n je počet objektov množiny so vzormi, a μ_{kl} je váha pridelená spojeniu prototypu $\mathbf{c}_k$ s vzorom $\mathbf{x}_l$, ktorá popisuje stupeň príslušnosti vzoru l do zhluku k. Keď c_k je najbližší (víťazný) prototyp k $\mathbf{x}_l$ v euklidovskej metrike, $\mu_{kl} = 1$, inak $\mu_{kl} = 0$. Odvodíme SCL minimalizovaním 1.29 z predpokladom, že váhy získavame použitím podmienky najbližšieho prototypu. Teda

$$E_l = \min_{1 \leqslant k \leqslant K} \|\mathbf{x}_l - c_k\|^2, \tag{1.31}$$

Na základe rovnice1.30 a použitím metódy gradient-descent, za predpokladu že $c_w = c_w(t)$ je víťazný prototyp pre $\mathbf{x} = \mathbf{x}_t$ dostaneme SCL ako:

$$\mathbf{c}_\omega(t+1) = \mathbf{c}_\omega(t) + \eta(t)[\mathbf{x}_t - \mathbf{c}_\omega(t)]$$
$$\mathbf{c}_j(t+1) = \mathbf{c}_j(t), i \neq \omega$$

kde $\eta(t)$ je zvolený tak, aby spĺňal Robbins-Monrove podmienky. Tento proces sa nazýva víťaz berie všetko. Tento algoritmus má dôležitú úlohu vo väčšine sieti nekontrolovaného učenia. Ak má každý zhluk svoj učiaci koeficient daný ako $\eta_i = \frac{1}{n_i}$, kde n_i je počet prvkov priradených $i-$tému zhluku, má algoritmus minimálny výstupný rozptyl.

Do podoby súťaživého učenia vieme previesť aj už uvedené zhlukovacie algoritmy. *K-means* zhlukovanie je blízko spojený z SCL a je to vlastne špeciálnym prípadom samo organizujúcich máp, ktoré sú veľmi populárnou metódou použitia neurónových sieti na zhlukovanie. Týmto algoritmom rozdelíme množinu n vstupných objektov do K oddelených podmnožín $\mathscr{C}_k$, kde každá podmnožina obsahuje n_k vstupných objektov minimalizáciou RSS

$$E(\mathbf{c}_1, ..., \mathbf{c}_K) = \frac{1}{n} \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in \mathscr{C}_k} \|\mathbf{x}_i - \mathbf{c}_k\|^2 \tag{1.32}$$

kde $\mathbf{c}_k$ je prototyp alebo ťažisko zhluku $\mathscr{C}_k$. Minimalizovaním E vzhľadom na $\mathbf{c}_k$, dostaneme optimálnu polohu prototypu $\mathbf{c}_k$ ako $\mathbf{c}_k = \frac{1}{n_k} \sum_{\mathbf{x}_i \in \mathscr{C}_k} \mathbf{x}_i$.

Takto vytvorený algoritmus môžeme aplikovať v hromadnej podobe, alebo v postupnej. Doteraz spomínaný algoritmus, bol o hromadnom vložení dátovej matice do algoritmu a nie je vhodný na súťaživé učenie. V prípade postupnej podoby, do algoritmu vkladáme dáta postupne

a takýto algoritmus je vhodný pre dáta ktoré sú získavané online. Do algoritmu môžme totiž vložiť úplne nové vzory. Tento algoritmus potom vieme zapísať nasledovne:

1. Náhodne vytvoríme K vektorov počiatočných prototypov zhlukov $m_1, ..., m_K \in \mathbf{R}^p$
2. Vložíme normalizovaný vzor $\mathbf{x} \in \mathbf{R}^p$
3. Vyberieme víťaza $E(\mathbf{c}_k)$ (teda $\mathbf{c}_k$ pre ktoré je hodnota funkcionálu E najmenšia).

4. Aktualizujeme víťazný prototyp pre $\mathbf{X}$:

$$\mathbf{c}_k(t+1) = \mathbf{c}_k(t) + \eta(t)(\mathbf{x} - \mathbf{c}_k(t)) \tag{1.33}$$

   kde η je učiaci koeficient.

5. Opakujeme kroky 2-4, kým nedosiahneme maximálny počet iterácii.

Koeficient učenia η teda určuje to ako sa vektor prototypov adaptuje k vstupnému vzoru a je priamo spojený s konvergenciou. Ak je $\eta = 0$ tak neprebieha žiadne učenie. Ak $\eta = 1$, tak je výsledok rýchlo učiaci sa prototyp.

1.6.1 Teória adaptívnych rezonančných váh - Adaptive resonance theory(ART)

Jedným zo základných problémov zhlukovania založenom na súťaživom učení je nestabilita. Dôvodom tejto nestability je 'plasticita' algoritmu, ktorá je potrebná na adaptovanie algoritmu pre nové vzory. ART ma schopnosť prispôsobiť sa novým dátam a pamätať si aj predchádzajúce tréningy a tým prekonáva problém stability-plasticity riešenia.

ART1 ART1 je najjednoduchší model s architektúrou ART, ktorý sa používa na zhlukovanie ľubovoľného množstva binárnych premenných vstupných vzorov (vektorov). Siete ART majú rekurentnú $J - K$ architektúru, ktorá sa líši od architektúry na Obr.5. Vstupná vrstva F1 sa nazýva porovnávacia vrstva a má J neurónov a výstupná vrstva F2 sa nazýva rozpoznávacia vrstva a má K neurónov. F1 a F2 sú plne prepojené oboma smermi a súťaženie prebieha vo vrstve F2 a jej súčasný stav sa nazýva krátkodoba pamäť (short term memory -STM). Neuróny z vrstvy F2, ktoré sú, už použité na reprezentovanie vstupného vzoru, sa nazývajú prepojené. Podobne, neprepojené neuróny nereprezentujú vstupný vzor. Zdola hore sú vrstvy prepojené váhovou maticu $W^{12} = \{w_{ij}^{12}\}$, pre spojenie i-tého neurónu z vrstvy F1 s j-tým neurónom vrstvy F2, a zhora dole sú prepojené váhovou maticou $W^{21} = \{w_{ji}^{21}\}$, ktorá sa tiež nazýva dl-hodobá pamäť (long term memory-LTM).

Vo vrstve F2 prebieha súťaženie medzi niektorými prepojenými neurónmi a jedným ne-prepojeným. Víťazný neurón pošle odhad váh naspäť do vrstvy F1. Odhad je porovnaný so vstupným vzorom. Vopred zvolený parameter opatrnosti ρ $(0 \leq \rho \leq 1)$ určí či sú odhad vzoru a vstupný vzor podobné. Ak sú si podobné a spĺňajú podmienku opatrnosti, adaptujeme obe váhy zároveň. Tento proces sa nazýva rezonancia. Ak podmienka nieje splnená, pošleme do vrstvy F2 signál, ktorý deaktivuje momentálny víťazný neurón. Čím vyššia je hodnota ρ o to menej nepresných vzorov bude tolerovaných a teda algoritmom vznikne viacej zhlukov.

Základné kroky ART1 zhrnieme nasledovne:

1. Inicializujeme váhy ako $w_{ij}^{12} = \xi/(\xi - 1 + p)$, kde p je počet dimenzii binárneho vstupného vzoru $\mathbf{x}$, $\xi > 1$ a $w_{ji}^{21} = 1$;
2. Do algoritmu vložíme nový vzor $\mathbf{x}$ a vypočítame presun z vrstvy F1 do vrstvy F2 ako

$$T_j = \sum_{i=1}^{d} w_{ij}^{12} x_i; \tag{1.34}$$

3. Vybraním víťazného neurónu J pravidlom víťaz berie všetko, aktivujeme vrstvu F2.

$$T_j = \max_j \{T_j\};$$

4. Porovnáme odhad z vrstvy F2 z vstupným vzorom. Ak

$$\rho \leq \frac{|\mathbf{x} \cap \mathbf{W}_J^{21}|}{|\mathbf{x}|},$$

kde $\cap$ je logická spojka AND, nasleduje krok 5, ak táto podmienka nie ej splnená tak pokračujeme krokom 6.

5. Aktualizujeme príslušné váhy pre daný aktívny neurón ako :

$$\mathbf{W}_j^{21}(new) = \mathbf{x} \cap \mathbf{W}_j^{21}(old), \tag{1.35}$$

a

$$\mathbf{W}_j^{12}(new) = \frac{\xi \mathbf{W}_j^{21}(new)}{\xi - 1 + |\mathbf{W}_j^{21}(new)|}. \tag{1.36}$$

Ak J je neprepojený neurón, vytvoríme nový neprepojený neurón s počiatočnými parametrami ako sú stanovené v kroku 1;

6. Pošleme do vrstvy F2 signál, ktorý deaktivuje momentálny víťazný neurón a vrátime sa do kroku 3;
7. Opakujeme kroky 2-7, dokým nie sú všetku vzory **x** priradené do zhlukov.

Fuzzy ART Fuzzy ART je ART sieť, ktorá je schopná naučiť sa rozpoznať stabilné zhluky pre binárne aj reálne vstupné vzory. Fuzzy ART má podobnú architektúru ako ART1 a používa operátory fuzzy množín na nahradenie binárnych operátorov, aby mohol pracovať aj s reálnymi dátovými množinami. Fuzzy ART popíšeme zvýraznením hlavných rozdielov medzi algoritmom fuzzy ART a ART1.

- Predspracovanie dát. Každý prvok z p-dimenzionálneho vstupného vzoru $\mathbf{x} = \{x_1, .., x_p\}$ musí byť v intervale $\langle 0, 1 \rangle$, a musí byť normalizovaný.
- Váhové matice ktoré spájajú vrstvy F1 a F2 inicializujeme ako v sieti ART1. Váhy neprepojeného neurónu nastavíme ako jedna.
- Výber zhluku. Po vložení vstupného vzoru, neuróny vrstvy F2 súťažia výpočtom funkcie výberu zhluku, ktorá je definovaná ako :

$$T_j = \frac{|\mathbf{x} \wedge \mathbf{w}_j|}{\alpha + |\mathbf{w}_j|}, \tag{1.37}$$

kde $\wedge$ je fuzzy logická spojka AND, a $\alpha > 0$ je parameter, ktorý používame na vybratie prototypu, keď sú viaceré ako jeden prototypový vektor fuzzy podmnožinou vstupného vzoru. Parameter α je závislý na parametre opatrnosti ρ, α by malo byť klesajúce ak ρ je klesajúce. Podobne ako v ART1, neurón J je aktivovaný pravidlom víťaz berie všetko

$$T_j = \max_j \{T_j\}.$$

- Priradenie zhluku. Funkcia výberu zhluku víťazného neorónu je potom otestovaná parametrom opatrnosti. Ak

$$\rho \leq \frac{|\mathbf{x} \wedge \mathbf{w}_J|}{|\mathbf{x}|},$$

nastane rezonancia. Inak je víťazný neurón zablokovaný a zvolíme a preskúmame nový neurón z vrstvy F2, ktorý porovnáme s parametrom opatrnosti. Tento proces opakujeme dokiaľ nie je podmienka splnená.

- Váhový vektor víťazného neurónu, ktorý spĺňa podmienku opatrnosti je aktualizovaný použitím učiaceho pravidla :

$$\mathbf{w}_J(new) = \beta(\mathbf{x} \wedge \mathbf{w}_J(old)) + (1 - \beta)\mathbf{w}_J(old), \tag{1.38}$$

kde je $\beta = \langle 0, 1)$ je učiaci parameter, $\beta < 1$, aby sme sa vyhli zbytočnému vytváraniu zhlukov a zhluky neboli zbytočne náchylné na šum v dátach.

1.7 Regresná analýza

V štatistickom modelovaní je pod pojmom regresia chápané hľadanie vzťahov medzi nezávislými premennými a premennými, ktoré na nich závisia. Regresná analýza je jedným z najpopulárnejších spôsobov modelovania, najmä vďaka jej relatívnej jednoduchosti a širokej škále aplikovanosti. Cieľom tejto diplomovej práce nie je detailný prehľad regresnej analýzy, preto v prípade potreby podrobnejšieho popisu odporúčame čitateľovi nahliadnuť do zdrojov, ktoré boli použité pri tvorbe tejto podkapitoly [1], [2] [19], [10], a [39].

1.7.1 Lineárna regresia

, Regresný model je pre náhodné veličiny $Y_1, ..., Y_n$ a maticu dát 1.1,typu $n \times p$, kde $p < n$. Predpokladajme že pre náhodný vektor $\mathbf{Y} = (Y_1, ..., Y_n)'$ platí

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{1.39}$$

kde $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ je vektor neznámych parametrov a $\mathbf{e} = (e_1, ..., e_n)'$ je náhodný vektor, ktorý spĺňa podmienku $E[\mathbf{e}] = \mathbf{0}$, $var[\mathbf{e}] = \sigma^2\mathbf{I}$, kde σ^2 je tiež neznámy parameter.

Pre takýto regresný model požadujeme, aby matica $\mathbf{X}$ nemala lineárne závisle stĺpce, čo znamená $h(\mathbf{X}) = p$, keďže predpokladáme že $p < n$. Z toho vyplýva, že matica $X'X$ je regulárna a existuje k nej inverzná matica.

Neznáme parametre $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$, ktoré sú presne dané rovnicou 1.27 však nie je možné nájsť bez toho aby sme ich počítali špeciálne pre každé $\mathbf{X}$ a $\mathbf{Y}$. Nahradíme ich odhadom $\mathbf{b} = (b_1, ..., b_p)'$, ktorý získame metódou najmenších štvorcov, teda minimalizáciou výrazu $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$. Tieto odhady počítame nasledovným vzorcom

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{1.40}$$

dôkaz refandel.
Hodnotu $\hat{\mathbf{Y}}$, ktorá predstavuje odhad $\mathbf{Y}$ získaný modelom, získame ako $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Maticu $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ zvykneme nazývať *hat matrix* a zapisovať $\mathbf{H}$. Táto matica je preimetnutím vektoru $\mathbf{Y}$ do regresného priestoru.

Rovnicu lineárneho modelu, ktorá obsahuje absolútny člen, môžme tiež zapísať ako :

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + e_i; i = 1...n.$$

Čo môže byť prepísané ako :

$$\mathbf{Y} = \mathbf{X}^*\boldsymbol{\beta}^* + \mathbf{e},$$

kde $\mathbf{X}^* = (\mathbf{1}_n, \mathbf{X})$ (na začiatok dát sme pridali stĺpec jednotiek).

1.7.2 Kvadratická regresia

Lineárnym regresným modelom je možné vyjadriť aj iný ako lineárny vzťah medzi závislou premennou $\mathbf{Y}$ a nezávislými premennými $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p)$. Zložitejšie vzťahy môžu byť modelované zahrnutím vhodných členov vyšších mocnín do vstupnej matice $\mathbf{X}$.

Príklad 1.11. Pre kvadratickú závislosť pre dvoch nezávislých premenných má teda regresná rovnica tvar:

$$Y_i = \beta_0 + \beta 1 x_{i1} + \beta 2 x_{i2} + \beta 3 x_{i1}^2 + \beta 4 x_{i2}^2 + \beta 5 x_{i1} x_{i2} + \epsilon_i, i = 1, ..., n$$

Stále hovoríme o lineárnom regresnom modely.

1.7.3 Hodnotenie modelu

Pre takto vytvorený model je vektor rezíduí, ktorý porovnáva pozorované hodnoty závislej premennej s odhadom ich stredných hodnôt, daný vzťahom $\mathbf{u} = \mathbf{Y} - \hat{\mathbf{Y}}$. Súčet štvorcov rezíduí je definovaný ako

$$RSS = d_{ES}(\mathbf{u}) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

Rozptyl rezíduí zavedieme ako $s^2 = RSS/n - p$. Variácia v závislej premennej je daná ako $ns_{YY} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$ a variácia vysvetlená regresným modelom $\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$. Medzi takto popísanými variáciami platí nasledujúci vzťah:

$$\sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

totálna variácia = variácia vysvetlená regresným modelom + súčet štvorcov rezíduí.

Koeficient determinácie r^2 nám podáva informáciu o tom, aká časť variácie je vysvetlená lineárnou závislosťou medzi premennými.Tento koeficient je štvorec mnohonásobnej korelácie medzi $\mathbf{X}$ a $\mathbf{Y}$.

$$r^2 = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}. \tag{1.41}$$

V extrémnych prípadoch, keď je celková variácia závislej premennej celkom vysvetlená modelom je $r^2 = 1$, druhý extrém je $r^2 = 0$, keď model nevysvetľuje žiadnu variáciu. môžme zapísať aj ako

$$r^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}. \tag{1.42}$$

Koeficient determinácie je ovplyvnený množstvom nezávislých premenných (regresorov). Pre danú vzorku s n objektami bude hodnota r^2 stúpať pridávaním regresorov do lineárneho modelu. Preto sa môže stať, že hodnota r^2 bude stúpať, aj keď budeme do modelu pridávať nevhodné regresory. Preto pre p regresorov ($p + 1$ premenných) zavedieme aj korigovaný (adjusted) koeficient determinácie ako

$$r_{adj}^2 = r^2 - \frac{k(1 - r^2)}{n - (k + 1)} \tag{1.43}$$

2 Modelovanie kvality ovzdušia

V zhlukovej analýze budeme aplikovať na údaje, ktoré súvisia s modelovaním kvality ovzdušia. V tejto kapitole stručne popíšeme rôzne druhy matematických modelov, ktoré sa používajú na riešenie rôznych problémov v oblasti kvality ovzdušia a aké výstupy tieto modely poskytujú.

Matematické modelovanie kvality ovzdušia je dôležitou skupinou metód, ktoré dopĺňajú monitorovanie kvality ovzdušia. Keďže monitoring kvality ovzdušia poskytuje informácie o kvalite ovzdušia iba v mieste monitorovacej stanice, modelovanie kvality ovzdušia sa používa v prípade, že je potrebné zhodnotiť celoplošné rozloženie koncentrácií znečisťujúcich látok, ďalej sa používa sa na predpoveď kvality ovzdušia, alebo na určenie podielu zdrojov na znečistení ovzdušia. Ďalšou veľkou oblasťou, kde sa uplatňujú postupy matematického modelovania, je určenie účinnosti opatrení na zlepšenie kvality ovzdušia. V ďalšom texte v krátkosti popíšeme druhy matematických modelov a príklady uplatnenia metódy FCM v problematike kvality ovzdušia.

V kapitole 2.1. popíšeme procesy, ktoré ovplyvňujú znečisťujúce látky v atmosfére a v kapitole 2.2 sa budeme venovať modelom kvality ovzdušia, keďže práve modely kvality ovzdušia poskytujú vstupné údaje pre túto prácu.

2.1 Znečisťujúce látky v atmosfére

Znečisťujúce látky sa dostávajú do ovzdušia z rôznych prírodných aj antropogénnych zdrojov znečisťovania ovzdušia a na ich ďalší osud vplývajú fyzikálne aj chemické procesy prebiehajúce v atmosfére. Niektoré znečisťujúce látky vznikajú v atmosfére pri chemických reakciách (napr. troposférický ozón, sekundárne častice, ako napr. sírany, dusičňany, atď.) alebo kondenzáciou horúcich pár. Chemickými reakciami sa znečisťujúce látky môžu meniť na menej, či viac toxické (vznik síranov, či nitro-polycyklických aromatických uhľovodíkov). Medzi fyzikálne procesy, ktoré ovplyvňujú koncentrácie znečisťujúcich látok v atmosfére patrí difúzia (rozptyl), advekcia (horizontálny prenos), konvekcia (vertikálny prenos), fázové prechody (zmeny skupenstva), suchá, mokrá a skrytá depozícia. (Suchá depozícia je sedimentácia znečisťujúcich látok účinkom gravitačnej sily, mokrá depozícia je vymývanie atmosférickými zrážkami a skrytá depozícia je záchyt na rôznych povrchoch. Prostredníctvom depozície sa znečisťujúce látky dostávajú z ovzdušia do ďalších zložiek životného prostredia - do vody, pôdy, sedimentov, následne sa môžu stať súčasťou bioty a potravinových reťazcov.

Antropogénnymi zdrojmi znečisťovania ovzdušia sú priemyselné zdroje, systémová energetika a výroba tepla,vykurovanie domácností, doprava (najmä cestná, ale tiež lodná doprava), poľnohospodárstvo, nakladanie s odpadom (skládkovanie a spaľovanie) a iné. Prírodné zdroje znečisťovania ovzdušia sú veterná erózia, vulkanická činnosť,vegetácia (emisie prchavých organických látok, pele), morská triešť (morská soľ a súvisiace minerálne látky, rozpustený organický uhlík), činnosť pôdnych bakterií, atď.

V našej analýze sa budeme zaoberať dvomi základnými znečisťujúcimi látkami - oxidom siričitým SO_2 a prachovými časticami PM_{10}.

Zdrojom oxidu siričitého na území Slovenska sú najmä priemyselné procesy - hlavne metalurgia, vrátane výroby koksu - a energetika [30]. Zdrojom PM_{10} môže byť aj vykurovanie domácností uhlím, čo sa prejavuje najmä na území Poľska v Malopoľskom a Sliezskom vojvodstve. Na území Slovenska sa uhlie na vykurovanie domácností používa v menšej miere, vzhľadom k dostupnosti palivového dreva.

Pod označením PM_{10} rozumieme prachové častice a drobné kvapôčky s aerodynamickým priemerom menším než 10 mikrometrov. Zdrojom PM_{10} je najmä vykurovanie domácností, cestná doprava, stavebné a búracie práce a poľnohospodárstvo. Emisie priemyselných zdro-

jov poklesli (vďaka legislatívnym opatreniam v posledných desaťročiach dvadsiateho storočia), na Slovensku sa prejavuje ich vplyv najmä v okolí metalurgického komplex, ostatné priemyselné zdroje sa prejavia vo svojom okolí v závislosti od meteorologických podmienok (zadymovanie). Vo všeobecnosti prispievajú veľké priemyselné a energetické zdroje skôr k vysokým pozaďovým koncentráciám, pretože znečisťujúce látky emitované z vysokých komínov sa v atmosfére za normálnych podmienok efektívne rozptyľujú. Stredná doba zotrvania SO_2 v troposfére je dva dni [38]. SO_2, ktorý je emitovaný do ovzdušia, sa môže oxidáciou meniť na sírany, alebo sa rozpustí vo vode na H_2SO_4 a opustí atmosféru cestou mokrej depozície, pričom podieľa na znížení pH atmosférických zrážok. Ďalším procesom odbúravanie SO_2 z atmosféry je suchá depozícia. Prachové častice PM_{10} sú naopak viac ovplyvnené chemickými procesmi v atmosfére, keďže až tretina z nich v atmosfére vzniká z plynných prekurzorov. Jemné prachové častice sú schopné diaľkového prenosu na stovky až tisíce kilometrov.

2.2 Modelovanie kvality ovzdušia

Na modelovanie kvality ovzdušia sa používa množstvo rozličných modelov, z ktorých každý je založený na istých zjednodušujúcich predpokladoch, ktoré určujú oblasť jeho použitia. Dôležitým faktorom, ktorý rozhoduje pri výbere modelu kvality ovzdušia pre riešenie danej úlohy je požadované priestorové rozlíšenie, či chemické a fyzikálne vlastnosti danej znečisťujúcej látky či skupiny látok. Procesy prebiehajúce v atmosfére sa týkajú rôznych priestorových mierok:

Mikroškálové -javy prebiehajúce na mierke 10 m - 100 m, napr. zadymovanie, či obtekanie budov alebo prúdenie v cestných kaňonoch. Na modelovanie šírenia znečisťujúcich látok v tejto mierke sa používajú CFD modely (Computational fluid dynamics).

Mezoškálové - javy prebiehajúce v mierkach po niekoľkých kilometrov až niekoľkých stoviek kilometrov (1000 m - 100 000 m) - napr. horsko-dolinná cirkulácia, bríza, atď.

Synoptické - javy prebiehajúce v mierkach niekoľkých stoviek až tisícov kilometrov, napr. presuny tlakových útvarov.

Globálne - javy v mierke viac než 5 tis. km. Napríklad pohyb systémov tlakových útvarov, stratosférický transport a oxidačné procesy prebiehajúce počas tohoto transportu. Globálne simulácie sa používajú aj pri štúdiu hemisférického transportu látok, ktoré zotrvávajú v ovzduší dlhú dobu (napríklad perzistentné organické látky) [25], [38].

2.3 Modely kvality ovzdušia

Jednou z možných kategorizácií modelov kvality ovzdušia je delenie na základe prístupu k riešeniu problému na modely deterministické, štatistické a fyzikálne. Deterministické modely spájajú príčinu s následkom. Medzi deterministické modely môžeme zaradiť gausovské, eulerovské a lagrangeaovské modely. Najčastejšie používanými štatistickými modelmi sú interpolačné techniky (napr. IDW (inverse distance weighting), kriging), neurónové siete, receptorové modely, regresné modely. Na modelovanie procesov v mezoškálovej mierkach až po globálne sa používajú eulerovské alebo lagrangeovské modely, prípadne gausovské rozptylové modely v závislosti od otázky, na ktorú sa modelovanie snaží odpovedať. Šírenie znečisťujúcich látok v atmosfére závisí od meteorologických podmienok, orografie (tvaru terénu) aj priestorového aj časového rozloženia emisií. Dôležité sú aj údaje o vlastnostiach povrchu (využitie krajiny - podiel priemyselnej, či urbánnej zástavby, druh vegetácie, atď.). Vstupom pre deterministický model kvality ovzdušia sú obvykle meteorologické údaje, (merania, alebo častejšie výstupy meteorologického modelu), ďalej sú potrebné podrobné údaje o emisných tokoch v celej výpočtovej doméne v požadovanom priestorovom a časovom rozlíšení (tieto sú obvykle

výstupmi emisných modelov).

Meteorologické modely riešia systémy parciálnych diferenciálnych rovníc, ktoré popisujú pohyb vzduchových hmôt v atmosfére. Ďalšie procesy, ako je mikrofyzika oblakov a zrážok, vplyv dlhovlnného a krátkovlnného žiarenia, vlastnosti povrchovej vrstvy, procesy v mestskej zástavbe sú parametrizované do formy empirických vzťahov. Tieto modely sú používané na predpoveď počasia, po reanalýze sa používajú na dlhodobé hodnotenie meteorologických situácií, aj na vytvorenie meteorologických údajov pre modely kvality ovzdušia.

Emisné modely pripravujú emisné vstupy pre modelovanie kvality ovzdušia v potrebnom priestorovom a časovom rozlíšení a v požadovanej projekcii. Vstupom emisných modelov sú obvykle celkové ročné emisie za celý štát, či menší územný celok v členení podľa druhu ľudskej aktivity, pri ktorej emisie vznikajú (napr. energetika, rozličné priemyselné činnosti, vykurovanie domácností doprava, atď. Emisný model potom na základe pomocných údajov (hustota obyvateľstva, tvar cestnej siete, teplota ovzdušia, atď.) rozdelí celkové emisie v čase a priestore. Výstupom emisného modelu je obvykle 4D pole priemerných hodinových emisných tokov - vstup pre konkrétny chemicko-transportný model kvality ovzdušia.
V nasledujúcej časti popíšeme bližšie niektoré najpoužívanejšie deterministické druhy modelov kvality ovzdušia a oblasť ich použitia.

2.3.1 Box model

Box model je najjednoduchší model materiálovej bilancie, popisuje koncentráciu znečisťujúcej látky v uzavretom priestore v tvare kvádra. Rozmery kvádra sú W(šírka po smere vetra) a L (dĺžka kolmá na smer vetra) v metroch, H môže dosahovať výšku premiešavania. Objem je definovaný ako $W \cdot L \cdot H[m^3]$.

Box model vychádza v najjednoduchšom prípade z nasledujúcich predpokladov:

- Lokálne emisie $Q[g/s]$, sú nezávislé na mieste a čase (spojité emisie). Vzťah medzi Q a mernou emisiou q na meter štvorcový je

$$Q = q \cdot (W \cdot L) \tag{2.1}$$

- Žiadne znečisťujúce látky sa nedostanú nad výšku premiešavania, emisie sa šíria len v smere vetra čiže nevytečú z objemu kvádra kolmo na smer vetra. Znečisťujúce látky nevstupujú do chemických reakcií a nepodliehajú sedimentácii.
- V boxe je koncentrácie znečisťujúcich látok homogénna.
- Smer a rýchlosť vetra sú konštantné $u[m/s]$.

Tieto predpoklady vedú k rovnovážnemu stavu a nulovej miere akumulácie.

Ak $\chi(t)[g/m^3]$ predstavuje koncentráciu znečisťujúcich látok ako funkciu času a χ_{in} je konštanta ktorá predstavuje koncentráciu emisii vo vstupujúcom vzduchu potom :
Vstupný tok $L \cdot H \cdot u \cdot \chi_{in}$
Výstupný tok $L \cdot H \cdot u \cdot \chi(t)$
Emisný tok Q
Rýchlosť úbytku 0
Zmena koncentrácie = $W \cdot L \cdot H \cdot \frac{d\chi(t)}{dt}$
Potom výsledná diferenciálna rovnica má tvar :

$$W \cdot L \cdot H \cdot \frac{d\chi(t)}{dt} = Q + L \cdot H \cdot u \cdot \chi_{in} - L \cdot H \cdot u \cdot \chi(t) \tag{2.2}$$

s riešením

$$\chi(t) = \frac{Q}{(L \cdot H \cdot u\cdot) \cdot (1 - e^{(-u \cdot \frac{t}{W})})} \tag{2.3}$$

Pre dlhší čas riešenie dosiahne ustálený stav $\chi(t) = \frac{Q}{(L \cdot H \cdot u \cdot)}$ čo zodpovedá nulovej zmene koncentrácie.

Niektoré box modely sú doplnené o základné chemické reakcie pre tú znečisťujúcu látku, ktorá je predmetom skúmania v konkrétnej riešenej úlohe. Výhodou box modelov je jednoduchosť a existencia analytického riešenia, nevýhodou je to, že mnohé zo jednodušení, ktoré sme použili sú nerealistické (napr. konštantná rýchlosť vetra, konštantný tok emisií, dokonalé premiešavanie v objeme boxu). Model nerozlišuje medzi veľkým počtom malých zdrojov emisií na povrchu od väčších zdrojov vo väčšej výške. Emisie všetkých zdrojov v boxe sa jednoducho sčítajú na získanie mernej emisie q, takže nemôžu popísať napr. tú skutočnosť, že zdroje vo väčšej výške spôsobujú v realite nižšie znečistenie pri povrchu. Box model preto poskytuje orientačné hodnoty použiteľné v idealizovanom prípade.

2.3.2 Gaussovské rozptylové modely

Gaussové rozptylové modely sú založené na nasledujúcich predpokladoch:

- Ustálené meteorologické podmienky (konštantná rýchlosť vetra v horizontálnom smere)
- Rýchlosť a smer vetra sa nemení s výškou.
- Rýchlosť vetra vo vertikálnom smere je zanedbateľná.
- Konštantný emisný tok.
- Vplyv chemických reakcií, mokrej aj suchej depozície je zanedbateľný.

Model rieši parciálnu diferenciálnu rovnicu, ktorá popisuje advekciu (horizontálny prenos v smere vetra) a turbulentnú difúziu.

$$\frac{\partial c}{\partial t} = D_y \frac{\partial^2 c}{\partial y^2} + D_z \frac{\partial^2 c}{\partial z^2} - u \frac{\partial c}{\partial x} \tag{2.4}$$

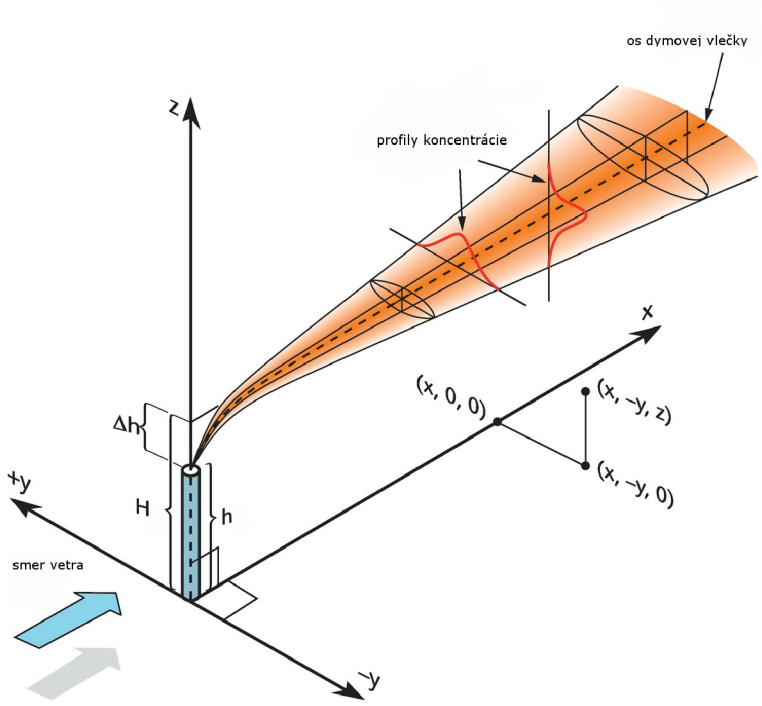
Posledný člen rovnice opisuje advekciu v smere osi x a prvé dva členy turbulentnú difúziu v smere kolmom na smer vetra, D_y a D_z sú koeficienty turbulentnej difúzie. Vplyv advekcie (prenos v smere vetra) obvykle značne prevyšuje vplyv turbulentnej difúzie. Pre bezvetrie rovnica nemá riešenie.

Počiatok súradnicovej sústavy je zvolený tak, aby bol v základni zdroja (päta komína), smer x-ovej osi je daný smerom vetra. Dymová vlečka opúšťa komín v bode $[0, 0, H]$ a stúpa do výšky $H = h + dh$, kde h je výška komína, dh je prevýšenie vlečky spôsobené počiatočnou hybnosťou spalín pri ústí komín a tepelným vznosom - rozdielom teploty spalín a teploty okolia a H je efektívna výška komína. Úloha má pri splnení spomínaných podmienok analytické riešenie - koncentrácia v smere kolmom na smer vetra v horizontálnom aj vertikálnom smere je popísaná Gaussovským rozdelením:

$$\chi(t) = \frac{Q}{2\pi \cdot u \cdot \sigma_y \cdot \sigma_y} e^{\frac{-y^2}{2\sigma_y^2}} \left[e^{\frac{-(H-z)^2}{2\sigma_z^2}} + e^{\frac{-(H+z)^2}{2\sigma_z^2}} \right] \tag{2.5}$$

V tejto rovnici $\sigma_y = a \cdot x^p$ je parameter disperzie v smere osi y vo vzdialenosti x od zdroja po vetre $\sigma_z = b \cdot x^q$ je parameter disperzie v smere osi z vo vzdialenosti x od zdroja. Konštanty a, b, p, q závisia na stabilite atmosféry [38].
Druhý exponenciálny člen v zátvorke popisuje úplný odraz dymovej vlečky od povrchu [38].

Gaussovské modely sa tiež používajú na výpočet príspevkov od viacerých zdrojov - postupne sa vypočítajú koncentrácie od všetkých bodových a líniových zdrojov (napríklad cestná komunikácia) a príspevky od jednotlivých zdrojov sa sčítajú.



Obr. 6: Reprezentácia dymovej vlečky v Gaussovskom modeli. Upravené podľa [15]

2.3.3 Lagrangeovské (trajektóriové) modely

Lagrangeovské modely sú založené na výpočte trajektórií ”balíčkov”vzduchu (znečisťujúcej látky) v závislosti od poľa vetra. Súradnicová sústava je pri lagrangeovských modeloch spojená s ”balíčkom”vzduchu na rozdiel od eulerovských modelov, ktoré počítajú koncentrácie v pravidelnej mriežke výpočtovej domény.

V porovnaní s Eulerovskými modelmi majú Lagrangeovské modely menšie nároky na výpočtové prostriedky a výpočet prebieha rýchlejšie, keďže výpočty chemických a fotochemických reakcii prebiehajú na menšom počte pohybujúcich sa buniek namiesto pevnej mriežky Eulerovského modelu. Niektoré z lagrangeovských modelov vychádzajú zo simulácie kontinuálnych emisií pomocou nespojitých emisií množstva jednotlivých balíčkov (”puffov”), ktorých poloha sa počíta v závislosti od diagnosticky vopred vypočítaného poľa vetra, pričom sa priebežne počíta aj ich rozptyl. Pre zvolené receptory a/alebo uzly mriežky výpočotovej domény sa potom spočíta príspevok všetkých puffov k výslednej koncentrácii v prízemnej vrstve. Lagrangeovské modely majú obvykle aj možnosť výpočtu spätných trajektórií (v tomto prípade sa simuluje pohyb opačne v čase, od receptora k zdroju). Príkladom často používaného lagrangeovského puff modelu je CALPUFF. Inými prípadom lagrangeovských modelov je Flexpart, ktorý simuluje emisie množstvom vypustených častíc. Príkladom hybridného lagrangeovsko-eulerovského modelu je Hysplit [32].

2.3.4 Eulerovské modely

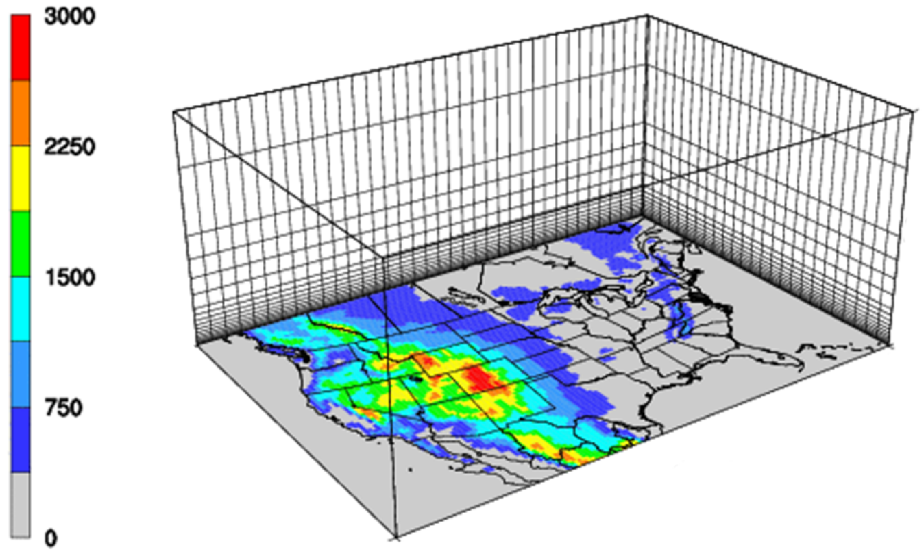
V eulerovských modeloch sa vzdušný priestor vo výpočtovej doméne rozdelí do pravidelnej mriežky (Obr. 7), ktorej bunky majú horizontálne priestorové rozlíšenie od 1 km do niekoľko desiatok km a vo vertikálnom smere sa vytvorí niekoľko vrstiev kopírujúcich terén. Pre každú bunku je splnená rovnica materiálovej bilancie.

V Eulerovských modeloch sa teda pre každú bunku pravidelnej 3D mriežky, v každom

časovom kroku rieši systém rovníc:

$$\frac{\partial c_i}{\partial t} = K_h \frac{\partial^2 c_i}{\partial x^2} + K_h \frac{\partial^2 c_i}{\partial y^2} + K_v \frac{\partial^2 c_i}{\partial z^2} - u \frac{\partial_i c}{\partial x} - v \frac{\partial c_i}{\partial y} - w \frac{\partial c_i}{\partial z} + E_i - D_i - R_i \tag{2.6}$$

kde prvé tri členy na pravej strane rovnice predstavujú turbulentnú difúziu, ďalšie potom advekciu v horizontálnom a vertikálnom smere, emisie (E_i), depozíciu D_i a chemické reakcie R_i pre i-tu chemickú látku zo súboru N látok popisujúci chemizmus v atmosfére. K_h a K_v sú koeficienty turbulentnej difúzie v horizontálnom a vertikálnom smere, u, v sú horizontálne zložky vetra a w vertikálna zložka vetra. Na začiatku model predpokladá počiatočné rozde-

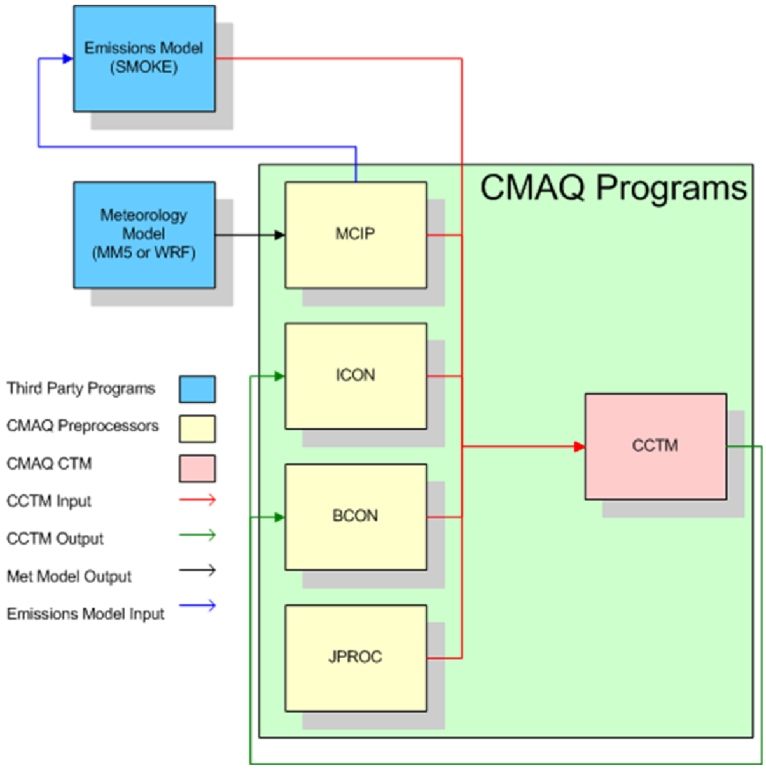


Obr. 7: Ilustrácia 3D mriežky výpočtovej domény v eulerovskom modeli.

lenie koncentrácií všetkých reaktantov. Následne sa pre každý časový krok (niekoľko minút) vypočíta zmena koncentrácií numerickým integrovaním základnej rovnice. Výpočet vyžaduje množstvo meteorologických dát, okrem iného údaje o smere a rýchlosti vetra v každej bunke, ktoré sú potrebné na výpočet toku z a do každej bunky cez jej hranice. CTM tiež zahŕňa výpočet chemických transformácii počas každého časového kroku. Výstupom sú obvykle 4D polia koncentrácií a hodnoty suchej a mokrej depozície znečisťujúcich látok.
Výhodou Eulerovských modelov je ich komplexnosť (zahŕňajú fyzikálne procesy - rozptyl aj prenos advekciou a konvekciou, suchú, mokrú aj skrytú depozíciu, chemické reakcie vo všetkých fázach, pričom pre väčšina procesov je dostupných niekoľko parametrizácií vhodných pre rozdielne podmienky). Preto sa tieto modely často označujú ako chemicko-transportné modely (CTM). Nevýhodou sú nároky na vstupné údaje, najmä emisie v jemnom časovom a priestorovom rozlíšení a vysoké nároky na výpočtové kapacity.

Najčastejšie používanými chemicko-transportnými modelmi sú CMAQ [5], CAM$_x$ [13], CHIMERE [23], EMEP model [34] a ďalšie.

Chemicko-transportný model CMAQ (The Community Multiscale Air Quality Modeling System) je open-source eulerovský model kvality ovzdušia, ktorý vyvíja agentúra US EPA (United States Environmental Protection Agency). CMAQ je súčasťou modelového systému, ktorý zahŕňa ešte meteorologický model a emisný model (Obr. 8). Tieto sa starajú o meteorologické a emisné vstupy do CMAQu v požadovanom priestorovom a časovom rozlíšení. CMAQ



.

Obr. 8: Bloková schéma modelu CMAQ

pozostáva z viacerých modulov - samostatných programov, ktoré pripravujú počiatočné a okrajové podmienky (ICON, BCON) a vypočítajú parametre fotochemických reakcií (JPROC) a interpolujú a dopočítavajú meteorologické polia (MCIP) ako vstupy pre samotné jadro modelu (CCTM).

Emisný model (SMOKE, emPy, FUME) je samostatným externým modelom, ktorý generuje hodinové emisné toky v pravidelnej mriežke v požadovanom rozlíšení pre CMAQ. Vstupom pre emisný model sú ročné emisie v členení podľa rôznych sektorov, čo zahŕňa mobilné, plošné aj bodové zdroje. Tieto emisie model rozdelí do mriežky podľa pomocných polí (údaje o využití krajiny, hustota osídlenia a pod.) a priradí im časový profil (ktorý sa napr. pri sektore vykurovania domácností odvíja od teploty vzduchu). Emisný model ďalej môže počítať tepelný vznos dymovej vlečky (v novej verzii CMAQu je táto funkcia zahrnutá)

Emisný model počíta aj biogénne emisie (prchavé organické látky, ktoré uvoľňuje do ovzdušia vegetácia). Keďže meteorologické podmienky ovplyvňujú aj šírenie dymovej vlečky a biogénne emisie, SMOKE používa dáta z meteorologického preprocesoru (MCIP).

Meteorologický preprocesor (MCIP) spracováva dáta z externého meteorologického modelu (WRF, Aladin alebo iného) do formátu, aký očakáva CMAQ. Súčasťou je interpolácia meteorologických polí do požadovaného horizontálneho rozlíšenia, orezanie výpočtovej domény a dopočítanie chýbajúcich parametrov.

Programy na výpočet počiatočných podmienok (ICON) a hraničných podmienok (BCON) poskytujú polia koncentrácií pre jednotlivé chemické prvky na začiatku simulácie a pre hranice modelovanej domény.Táto funkcia používa dáta z predchádzajúcich simulácií modelu CMAQ alebo z vertikálnych profilov troposféry pre čistý vzduch.

Simulácia procesov v CMAQu CMAQ simuluje okrem vertikálneho a horizontálneho prenosu (konvekcia, advekcia), difúzie a depozície aj procesy v oblakoch a tvorbu atmosférického aerosólu. Oblaky majú priamy aj nepriamy vplyv na koncentrácie znečisťujúcich látok: priamo ovplyvňujú koncentrácie prostredníctvom chemických reakcii s vodou, vertikálneho zmiešavania a mokrej depozície a nepriamo ovplyvňujú mieru fotolýzy. Dôležitým procesom je vymývanie zrážkami (mokrá depozícia).

Výstupné údaje modelu CMAQ

Výstupom CMAQu sú polia priemerných hodinových koncentrácií, dohľadnosti, suchej a mokrej depozície v pravidelnej mriežke so zvoleným horizontálnym rozlíšením v prízemnej vrstve a v zadaných vertikálnych hladinách. Vstupné aj výstupné súbory sú v netCDF formáte.

2.3.5 Hranice použiteľnosti jednotlivých druhov matematických modelov kvality ovzdušia

Eulerovské modely sú použiteľné pre simulácie, kde je postačujúce priestorové rozlíšenie niekoľko kilometrov, obvykle sa ako minimálne rozlíšenie používa 500m až 1km. v závislosti od komplexnosti terénu [31]. Často je výpočet nastavený pre niekoľko vnorených domén, pričom vnútorné rozlíšenie vnútornej domény je v pomere 1:3 k rozlíšeniu vonkajšej, pričom vonkajšia (materská) doména poskytuje vnútornej okrajové podmienky. Výstupy CTM modelov obvykle nemajú dostatočné rozlíšenie pre úlohy týkajúce sa mestského prostredia, ich výhodou je, že sú schopné popísať procesy prebiehajúce v regionálnej mierke. V urbánnej zástavbe môžu byť použité gaussovské modely, so započítaním drsnosti povrchu a vplyvu cestných kaňonov, pričom výstupy CTM modelu môžu byť použité ako okrajové podmienky. Pre simuláciu rozptylu znečisťujúcich látok po priemyselných haváriách sa môžu použiť pred-vypočítané výstupy z mikroškálových CFD (Computational fluid dynamics) modelov, ktoré možno použiť pre priestorové rozlíšenie niekoľkých metrov [29].

2.3.6 Použitie zhlukovej analýzy v oblasti kvality ovzdušia

Keďže každý z modelov kvality ovzdušia má svoje obmedzenia, pri komplexnejších úlohách sa používa viacero modelov spojených do logickej reťaze, ktorej súčasťou veľmi často bývajú aj štatistické metódy. Príkladom použitia štatistických metód je predpoveď kvality ovzdušia pomocou umelých neurónových sietí, ktorá má často lepšiu úspešnosť ako deterministické modely [6] Zhluková analýza si v problematike súvisiacej s kvalitou ovzdušia našla uplatnenie vo viacerých oblastiach - napr. pri spracovaní údajov z nízko rozpočtových senzorov [20], zlepšenie úspešnosti predpovede kvality ovzdušia, či výpočet indexov kvality ovzdušia pre informovanie verejnosti (ref.)

V oblasti výskumu modelovania kvality ovzdušia sú používané aj fuzzy množiny, zhlukovacie metódy ktoré ich používajú a fuzzy inference systémy (FIS). Algoritmus fuzzy c-means bol použitý, na vyhodnotenie kvality siete monitorovacích staníc PM_{10} a SO_2 a nájdenie staníc, ktoré je možné na základe podobných podmienok a nameraných výsledkov zredukovať [9], ďalej bol použitý na nájdenie miest, v ktorých najmä kvôli dopravným emisiám sa zhoršuje kvalita ovzdušia a je v nich ešte možné tomuto zhoršeniu zabrániť [8]. V modelovaní kvality ovzdušia si však väčšie uplatnenie našiel FIS ktorý je často skúmaný ako možný model, ktorý použiť na výpočet indexu kvality ovzdušia[12] a tiež bol použitý aj pri analýze dopadu kvality ovzdušia na verejné zdravie spolu s FCM algoritmom[17].

3 Aplikácia

V tejto kapitole sa budeme venovať aplikácii zhlukovej analýzy na reálne dáta. Budeme analyzovať vzťah emisií, meteorologických parametrov a údajov o mestskej zástavbe. Tieto údaje sú súčasťou vstupných údajov pre model CMAQ.

Použijeme emisie oxidu siričitého ($SO_2[mol/sek]$) a prachových častíc ($PM_{10}[g/s]$) z roku 2017. Spomínané znečisťujúce látky sme zvolili, na základe konzultácie s odborníkmi z SHMU práve, na základe ich vlastností a miery, akou závisia na meteorologických parametroch.

Výsledky zhlukovej analýzy použijeme na vytvorenie jednoduchého regresného modelu, ktorý bude zachytávať vzťah medzi výstupmi modelu CMAQ (koncentráciami SO_2, resp.PM_{10}) a vstupmi modelu (emisiami SO_2, resp.PM_{10}) a premennými použitými pri zhlukovej analýze (Tab.1).

Model CMAQ je náročný na prípravu vstupov, objem vstupných a výstupných dát aj výpočtový čas. Pre jediný deň je potrebných 219 premenných, v hodinových intervaloch na doméne s 18952 bunkami. To je 99 611 712 dátových bodov pre jediný deň, čo sú megabajty dát. Mi sme s takým množstvom dát nepracovali. Zvolili sme subdoménu a po konzultácii s pracovníkmi ústavu sme vybrali len niekoľko najdôležitejších veličín.

Nami vytvorený jednoduchý model by umožňoval rýchlejšie a flexibilnejšie výpočty koncentrácii znečisťujúcich látok. To však nie za účelom nahradenia modelu CMAQ, ale za účelom skôr predbežného odhadu koncentrácií - napr. pri simulovaní vplyvu rôznych emisných scenárov. Takýto model vytvoríme v nasledujúcich krokoch:

1. Korelačná analýza.
2. Úprava dát na zhlukovú analýzu
3. Zhluková analýza. Porovnanie dvoch metód zhlukovania:
   (a) K-means
   (b) fuzzy C-means
4. Vyhodnotenie výsledkov zhlukovania
5. Nájdenie regresných vzťahov medzi cmaqovskými vstupmi a výstupmi pre jednotlivé zhluky.
6. Porovnanie modelov

Tento postup sme použili zvlášť na dataset obsahujuci SO_2 a PM_{10}, pri zachovaní ostatných premenných, aplikovali sme oba na kratší časový úsek bez zrážok a pre porovnanie na celý rok dát.

3.1 Zhluková analýza

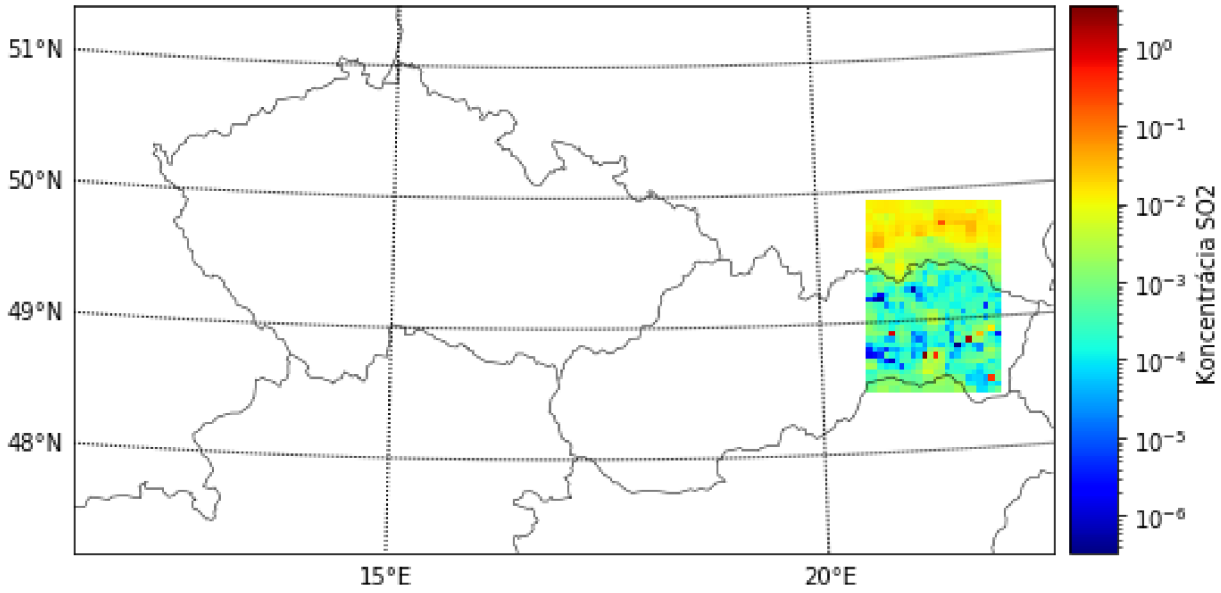
V tejto kapitole bližšie popíšeme proces a výsledky zhlukovania pre zvolené veličiny. Proces budeme podrobnejšie popisovať a vizualizovať pre emisie SO_2. Výsledky a proces pre zhlukovanie PM_{10} popíšeme slovne a vizualizácie zahrnieme do prílohy.

3.1.1 Popis dát

Na analýzu sme z pôvodnej výpočtovej domény CMAQu, ktorá pokrývala územie SR a ČR (Obr. 9) zvolili subdoménu, ktorá zahŕňa východné Slovensku a časť Malopoľského vojvodstva. Subdoména bola vybraná tak, aby obsahovala rôzne zdroje emisií SO_2 - vykurovanie domácností s použitím uhlia s vysokým obsahom síry (oblasti v Poľsku) a priemyselný zdroj - metalurgický komplex US Steel pri Košiciach. Zvolili sme 11 za sebou idúcich dní bezzrážkového obdobia, počas ktorých nedochádzalo k mokrej depozícii, aby sme na úvod analyzovali čo naj-

jednoduchšiu situáciu. Ako toto obdobie sme za pomoci meteorológov vybrali dni od 15. do 26. januára. Týmto vytvoríme akúsi rozšírenú verziu podmienok box modelu (2.3.1), ktorý závisí najmä na smere a rýchlosti vetra.

Okrem vstupných údajov o emisiách sme do dátovej matice vložili základne údaje o mete-



Obr. 9: Subdoména

orológii, ktoré ovplyvňujú šírenie znečisťujúcich látok v atmosfére. Tieto veličiny sú uvedené a popísané v Tabuľke.1. Meteorologických parametrov, ktoré ovplyvňujú osud rozptyl a prenos znečisťujúcich látok je viac. Údaje ako množstvo zrážok a iné sme momentálne nezaradili do práce, aby sme vytvorili relatívne jednoduchý model a získali čo najprehľadnejšie výsledky na ďalšiu analýzu. Keďže sme pracovali s agregovanými veličinami denných priemerov, zahrnuli sme rýchlosť a nie smer vetra.

Nami zvolená subdoména je zobrazená farebne na Obr. 9 (čiernobiela časť obrázku je výpočtová doména CMAQu). Subdoména je pokrytá mriežkou, ktorá obsahuje 25 x 35 buniek, rozmer jednej bunky je daný horizontálnym rozlíšením, ktoré bolo použité pri simulácii CMAQu ($4,5km \times 4,5km$).

Použili sme 11 dní a 25 x 35 buniek, čím sme získali 9625 objektov pre zhlukovanie. Okrem údajoch o emisiách sme do dátovej matice vložili vybrané meteorologické a orografické veličiny. Základné popisné štatistiky použitých veličín sú uvedené v Tabuľke.1.

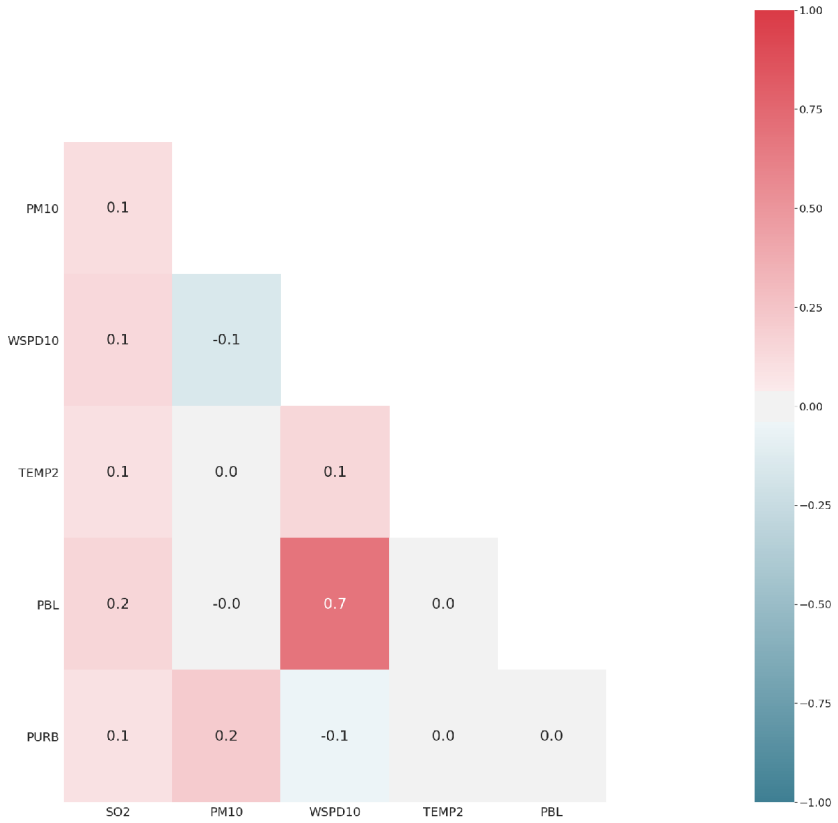
Tabuľka 1: Priemerné hodnoty a rozptyl dát

Veličina	$E(\cdot)$	$\sigma(\cdot)$	Popis veličiny $[jednotky]$
SO_2	$1,48 \cdot 10^{-2}$	$1,57 \cdot 10^{-1}$	emisie oxidu siričitého $[mol/sek]$
PM_{10}	0.16	0.528	častice s aerodynamickým priemerom menším ako 10 μm $[g/s]$
WSPD10	$2,92$	$1,52$	rýchlosť vetra vo výške $10m$ nad zemou $[m/s]$
TEMP2	267.67	$1,78$	teplota $[K]$
PBL	$158,49$	$80,51$	výška premiešavania $[m]$
PURB	0.28	2.15	percento urbánnej zástavby

3.1.2 Proces zhlukovej analýzy

Dáta sme spracovávali použitím programovacieho jazyka Python. Proces sme začali výpočtom popisných štatistík o dátovej matici a výpočtom korelačných koeficientov medzi jednotlivými nameranými veličinami použitím knižnice **pandas**. Vzhľadom na podmienku normality dát, ktorú naše dáta nespĺňajú, sme zvolili Spearmanov korelačný koeficient, ktorý zachytáva nelineárnu závislosť medzi premennými. Korelačné koeficienty sme umiestnili do korelačnej matice. Korelačnú maticu sme zobrazili použitím knižnice **seaborn** (Obr.10).

Z korelačnej matice vyčítame, že pre premennú PBL je korelácia s premennou WSPD10

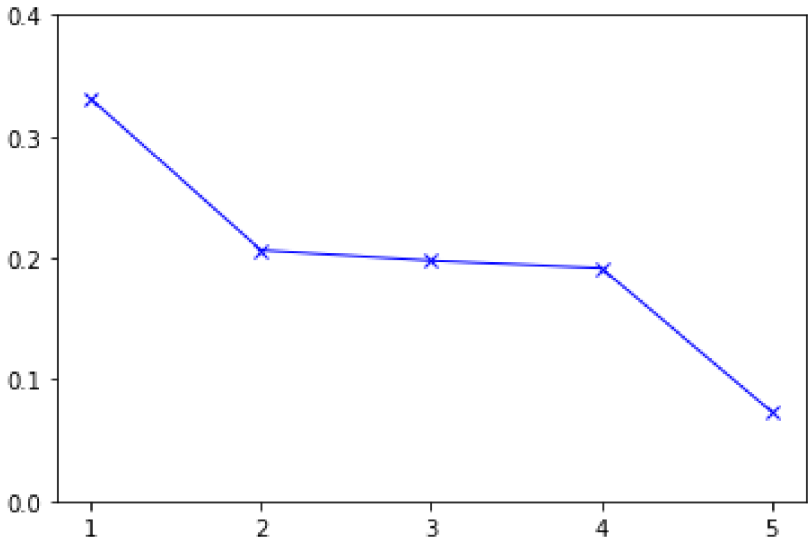


Obr. 10: Korelačná matica pre veličiny použité na zhlukovanie.

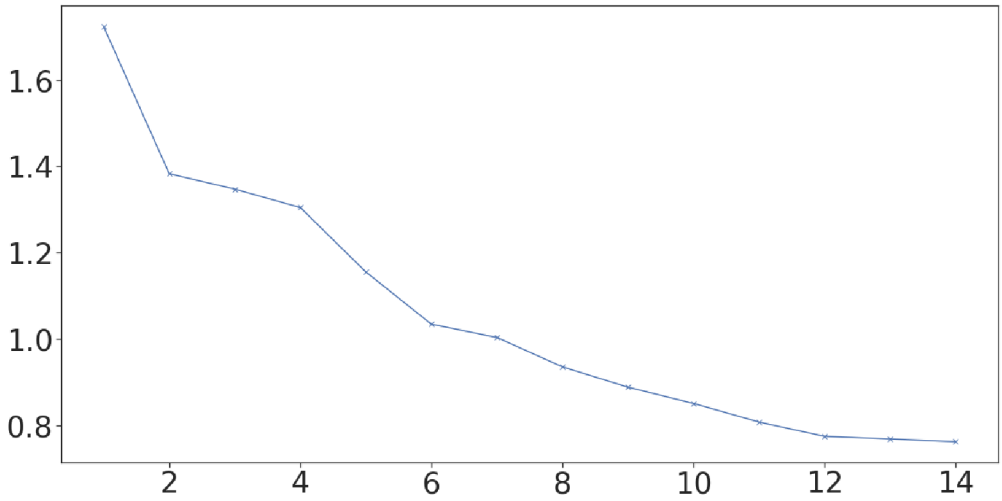
pomerne vysoká a to $0,7$. To znamená že aj keď pre ostatné dvojice premenných sú korelačné koeficienty nízke, je potrebné tieto dáta transformovať na nekorelované.
Z popisnej štatistiky týchto dát môžme porovnať stredné hodnoty a rozptyly premenných. Premenná PBL má strednú hodnotu $158,49$ metrov a rozptyl $80,51$ metrov, zatiaľ čo emisia SO_2, ktorú sme zvolili na pozorovanie má strednú hodnotu$1,48 \cdot 10^{-2}$ a rozptyl $1,57 \cdot 10^{-1}$. Ak by sme zhlukovali na takýchto dátach, podobnosť objektov by ovplyvňovala najmä ich hodnota premennej PBL a na premennej SO_2 by záviselo rádovo menej. Preto sme dáta pred ďalšími krokmi štandardizovali.

Na transformáciu dát na nekorelované sme použili analýzu hlavných komponent, ktorá sa v bežnej praxi používa aj na redukciu počtu premenných. Nové premenné sú zoradené podľa toho, koľko z celkového rozptylu dát vyjadrujú. Na zobrazení hodnôt vlastných čísiel na grafe na Obr.11, môžme pozorovať, že po takejto transformácii posledný piaty komponent popisuje $7,36\%$ rozptylu v dátach. Vzhľadom na to, že sme už na úvod, za pomoci odborníkov, vybrali čo najmenej premenných, ktoré ovplyvňujú šírenie emisií, nebudeme tento komponent eliminovať.

Na takto štandardizovaných a nekorelovaných dátach môžeme ďalej vykonať zhlukovú



Obr. 11: Vlastná čísla výsledných komponentov PCA. Na ose Y je hodnota vlastného čísla, na ose x sú jednotlivé komponenty

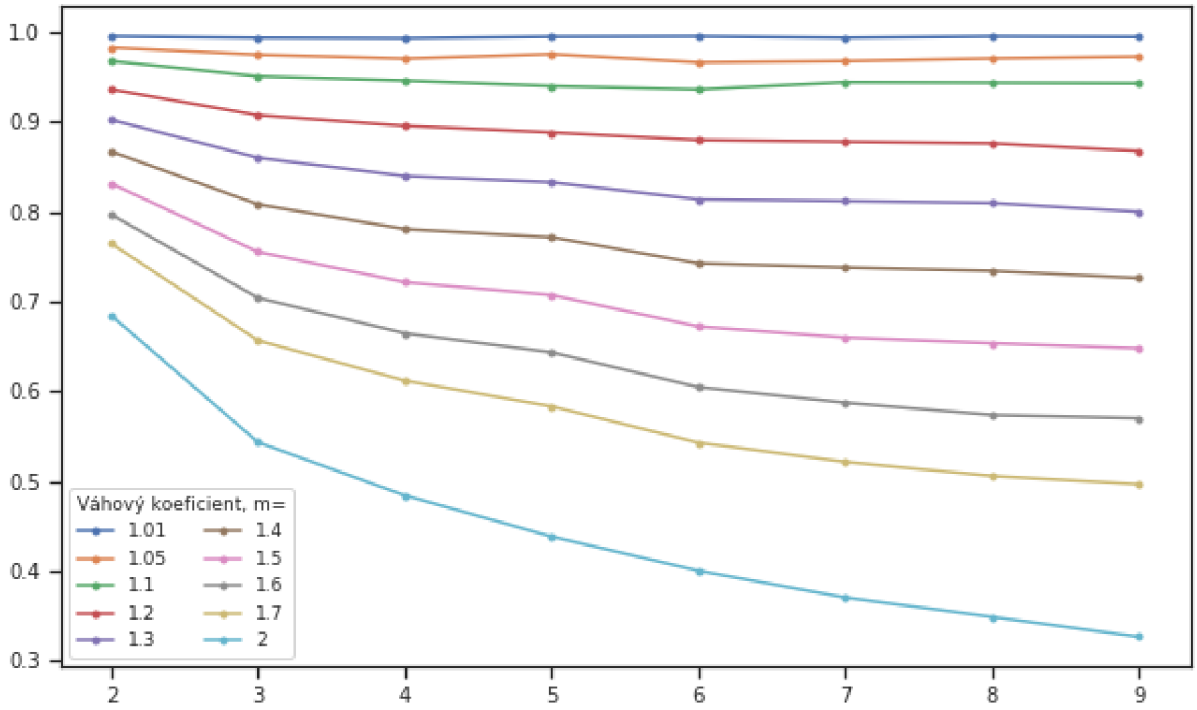


Obr. 12: Elbow Method aplikovaná na dátovú maticu s SO_2. Na ose Y je súčet štvorcov odchýliek objektov od ťažiska zhluku, do ktorého boli priradené. Na ose X je počet zhlukov.

analýzu. Ako prvé je potrebné zistiť, aký je optimálny počet zhlukov pre našu dátovú maticu. Použili sme lakťovú metódu, ktorej výsledok sme zobrazili do grafu na obr.12. Hneď prvý veľký skok je pri zmene počtu zhlukov z jedného zhluku na dva. Avšak hodnota funkcionálu 1.21 je stále relatívne vysoká. Ďalšia zmena v klesaní hodnôt funkciunálu je pri hodnote šiestich zhlukov. Dáta teda budeme zhlukovať s 6 zhlukmi.

Pre použitie FCM na analýzu dát bolo potrebné zvolenie váhového koeficientu m do funkcinálu 1.22 a nájdenie optimálneho počtu zhlukov pri použití FCM. Na tento účel sa v zdrojoch [33] používa fuzzy partition coeficient. Na grafe na Obr.13, ktorý je podobný ako pre lakťovú metódu, hľadáme počet zhlukov, pre ktorý je hodnota fpc vyššia ako pre iný počet zhlukov, pre rôzne hodnoty váhového koeficientu, ktoré sú farebne označené. Pre túto hodnotu, majú totiž zhluky čo najjasnejšie hranice, teda je v nich rozoznateľná štruktúra zhlukov.

Pri hľadaní optimálneho váhového koeficientu sme veľmi rýchlo spozorovali, že hodnoty



Obr. 13: Závislosť hodnoty FPC na počte zhlukov pre rôzne hodnoty váhového koeficientu

závislosti koeficientu fpc na množstve zhlukov sa pre hodnoty váhového exponentu blízke 2 blížia k hodnotám $1/K$, ktoré predstavujú pre tento koeficient minimálne hodnoty. Pre nízke hodnoty váhového koeficientu blízke 1, bude hodnota fpc vysoká, keďže nižšie hodnoty váhového koeficientu majú výsledné zhluky fuzzy c-means algoritmu hranice, ktoré sú takmer pevné.

Z grafu na Obr.13, môžeme vidieť, že hodnoty fpc výrazne nestúpnu pre žiaden počet zhlukov pre žiadnu hodnotu váhového koeficientu. Skúsili sme teda vypočítať minimálnu hodnotu váhoveho exponentu podľa [36], avšak naša dátová matice nespĺňa základné predpoklady výsledku tohto článku. Hodnota najväčšieho vlastného čísla, špeciálnej matice F_{U^*}, je väčšia ako $0,5$. Rozhodli sme sa preto porovnať dva možné scenáre, pre rôzne odporúčané minimálne hodnoty váhového exponentu. Zvolili sme minimum, ako je odporúčané v [4] a to $1,05$. Pre túto hodnotu m môžme vidieť malé zvýšenie hodnoty fpc pre 5 zhlukov, použijeme teda aj tento výsledok a porovnáme ho s minimálnym váhovým exponentom podľa Bezdeka[28] $1,5$, pre ktorý zvolíme rovnaký počet zhlukov ako pre k-means zhlukovanie, čo nám následne zjednoduší aj porovnávanie výsledkov. S takýmito parametrami bude mať fpc pre váhový exponent $1,5$ hodnotu $0,76$ a pre váhový exponent rovný $1,05$ bude hodnota fpc $0,96$.

Priebeh analýzy pre PM_{10} bol obdobný. Dátovú maticu sme štandardizovali a transformovali metódou PCA. Lakťovou metódou sme získali rovnaký optimálny počet zhlukov do algoritmu k-means ako pre dátovú maticu s emisiami SO_2. Jediný rozdiel bol vo výbere optimálneho počtu zhlukov pre algoritmus FCM. Počet zhlukov pre nižšiu hodnotu váhového exponentu, kde sme na základe drobného nárastu hodnoty FPC pre $m = 1,05$ zvolili ako optimálny počet 6 zhlukov. Pre PM_{10} teda budeme ďalej porovnávať tri rôzne možnosti, z ktorých každá má šesť zhlukov.

Následne sme na dátovej matici spravili zhlukovú analýzu so zvolenými parametrami v programovacom jazyku python. Použili sme algoritmus *k-means* ako je uvedený v kapitole 1.6.1 a *fuzzy c-means* algoritmus ako je uvedený v kapitole 1.6.2, použitím knižníc [40], [41], [42], [43] a [45] a na vizualizáciu sme použili knižnice [41] a [44]. Výsledky zhrnieme v ďalšej kapitole.

3.1.3 Výsledky zhlukovej analýzy

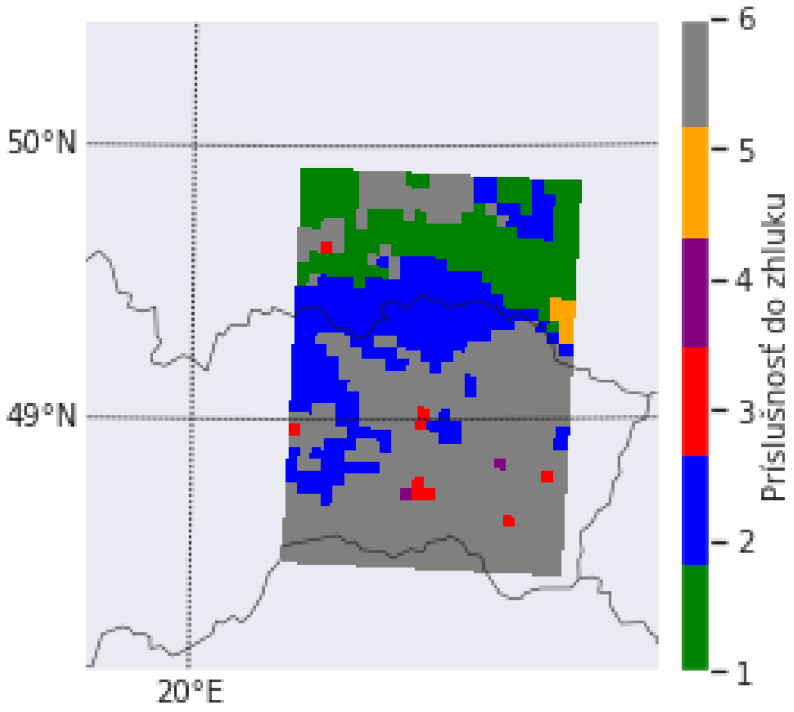
Mohutnosti výsledných zhlukov algoritmu k-means sú uvedené v Tabuľke 2. Algoritmus vytvoril nerovnomerné zhluky. Čo v mnohých praktických prípadoch môže to byť spôsobné zastavením algoritmu v lokálnom minime. Ale pri bližšom pohľade na mohutnosti zhlukov kmeans_3 a kmeans_5 a ich zobrazení na mape si môžme všimnúť, že mohutnosť týchto zhlukov je deliteľná 11, a teda ide o miesta zo špecifickými hodnotami vstupných veličín, ktoré tvoria v dátach

Tabuľka 2: Mohutnosti výsledných zhlukov algoritmu k-means

Zhluk	Mohutnosť zhluku
kmeans_1	2254
kmeans_2	2326
kmeans_3	99
kmeans_4	22
kmeans_5	1861
kmeans_6	3063

oddelené zhluky. Metódou k-means sme v dátach našli štruktúru, ktorá je nerovnomerná.

K výsledným zhlukom sme pridali informáciu o zemepisnej výške a zemepisnej dĺžke a následne sme ich pre jednotlivé dni zobrazili na mape na Obr.14.



Obr. 14: Zobrazenie výsledných zhlukov algoritmu k-means zhluk pre 15. deň januára v závislosti na zemepisnej šírke a zemepisnej dĺžke

Algoritmom fuzzy c-means sme získali maticu príslušností do zhlukov. Aby sme mohli výsledky algoritmu porovnať s výsledkami k-means, pre každý objekt sme zvolili zhluk do ktorého patrí, ako ten pre ktorý ma objekt najvyšší stupeň príslušnosti. Pre zhlukovanie s váhovým koeficientom $1,5$ sú výsledné mohutnosti šiestich zhlukov uvedené v Tabuľke 3. A na prvý

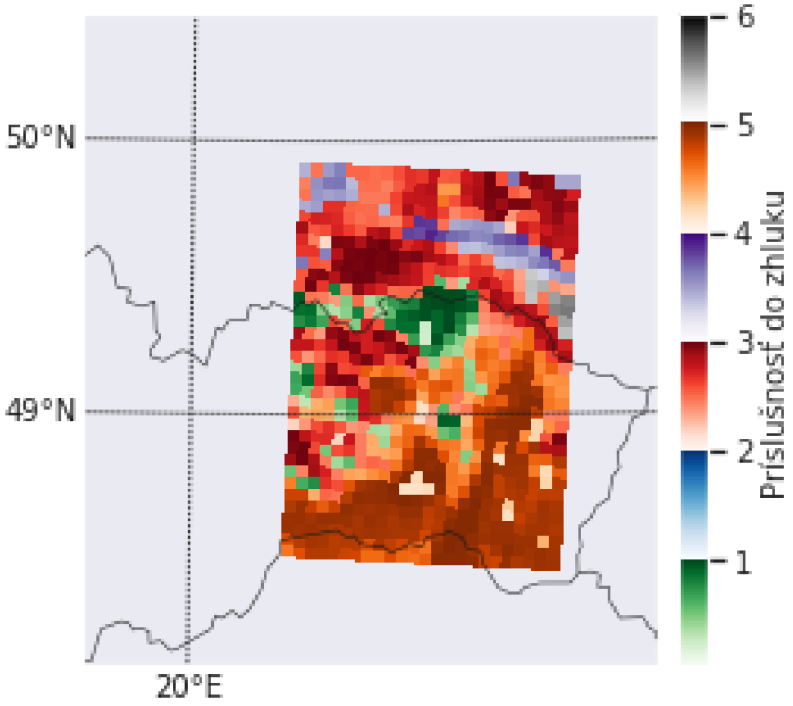
pohľad môžme vidieť, že pre fuzzy zhlukovanie sú zhluky rovnomernejšie rozdelené. To je v mnohých prípadoch žiadúce, môže to znamenať, že v našich dátach je rovnomerná štruktúra, alebo to, že algoritmus fuzzy c-means nebol schopný zachytiť štruktúru, čo mohlo viesť k tomu, že objekty boli takmer rovnomerne rozmiestnené do všetkých zhlukov. Vzhľadom na výsledky metódy k-means, ktorá je tiež závislá na počiatočnom rozklade, je možné že ide o druhú možnosť

Pri výslednej matici príslušnosti sme postupovali nasledovne. Pre každý objekt sme zvo-

Tabuľka 3: Mohutnosti výsledných zhlukov algoritmu fuzzy c-means s hodnotou váhového koeficientu $m = 1,5$

Zhluk	Mohutnosť zhluku
fuzzy_1	1508
fuzzy_2	871
fuzzy_3	1758
fuzzy_4	1683
fuzzy_5	2484
fuzzy_6	1321

lili zhluk do ktorého patrí, ako ten pre ktorý ma objekt najvyšší stupeň príslušnosti. V ďalšom kroku sme od čísla, ktoré klasifikuje zhluk, odčítali súčet príslušnosti objektu k ostatným zhlukom, teda to, čo objektu chýba na to, aby patril iba tomuto zhluku. K takto vytvorenému vektoru sme pridali informácie o zemepisnej šírke a zemepisnej dĺžke a zobrazili sme ho na mape na Obr.15 a Obr.16 pre jednotlivé dni. Získali sme tak zobrazenie nielen toho, pre aký zhluk má objekt najvyšší stupeň príslušnosti, ale aj konkrétnu informáciu o stupni príslušnosti samotnom (čím je objekt na mape zobrazený výraznejšie, tým viac do daného zhluku patrí).



Obr. 15: Zobrazenie výsledných zhlukov algoritmu fuzzy c-means s hodnotu váhového exponentu $m = 1,5$ pre 15. deň januára v závislosti na zemepisnej šírke a zemepisnej dĺžke

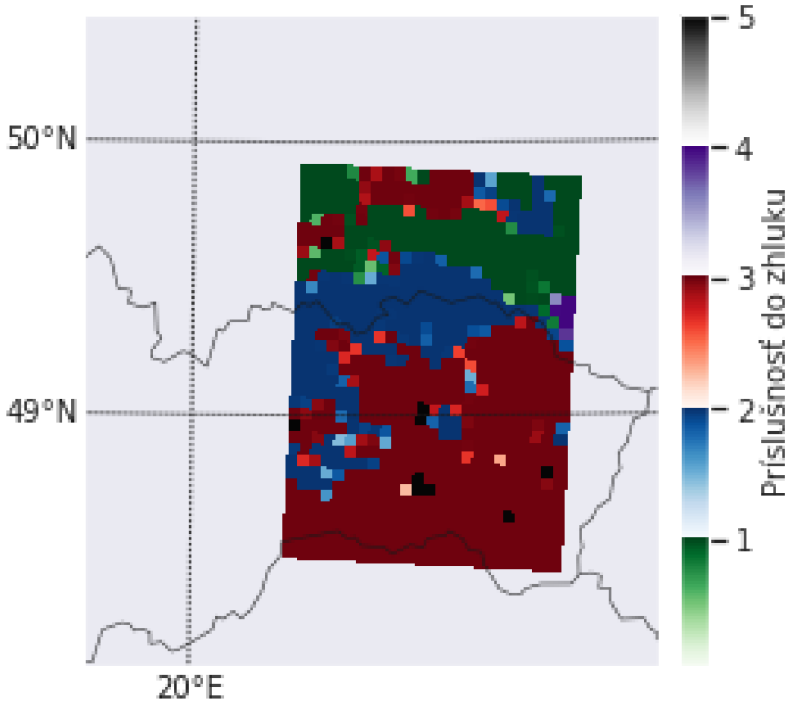
Pre zhlukovanie s piatimi zhlukmi a hodnotou váhového koeficientu $1,05$ sme tak 5 zhlukov s mohutnosťami uvedenými v Tabuľke 4. Mohutnosti týchto zhlukov sú podobnejšie výsledným zhlukom z algoritmu k-means.

Tabuľka 4: Mohutnosti výsledných zhlukov algoritmu fuzzy c-means s hodnotou váhového koeficientu $m = 1,05$

Zhluk	Mohutnosť zhluku
fuzzy_1	2253
fuzzy_2	2327
fuzzy_3	3079
fuzzy_4	1867
fuzzy_5	99

To isté môžme pozorovať aj z mapy na Obr.16, z ktorej môžme vidieť, že zobrazené zhluky sú tvarmi podobné výsledným zhlukom algoritmu k-means

Pre všetky tieto zobrazenia na Obr.14, Obr.15 a Obr.16 je potrebné si poznamenať, že



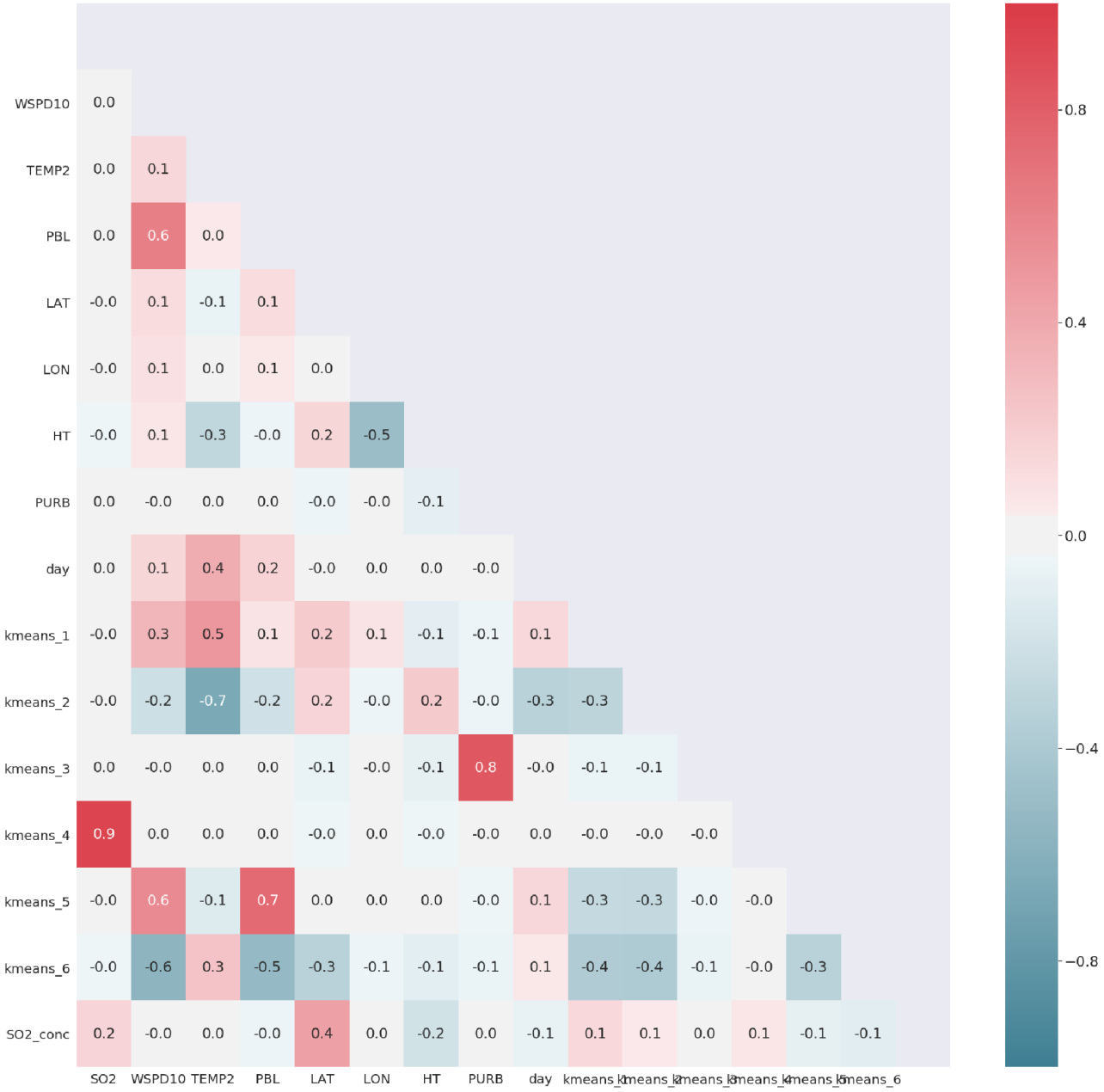
Obr. 16: Zobrazenie výsledných zhlukov algoritmu fuzzy c-means pre hodnotu váhového exponentu $m = 1,05$ pre 15. deň januára v závislosti na zemepisnej šírke a zemepisnej dĺžke

počiatočné ťažiská boli iniciované náhodne a čísla ktorými sú ”pomenované”jednotlivé zhluky sú závislé na tomto počiatočnom delení a preto nenesú žiadnu inú informáciu ako na identifikovanie jednotlivých zhlukov. Štruktúry, ktoré sú zobrazené na mapách, sú komplexné a je náročné porovnať ich s vizuálnym zobrazením použitých veličín na mape, ako sú zobrazené emisie na Obr. 9 a koncentrácie na Obr.20.

Na získanie kvantifikovaného porovnania výsledkov zhlukovania a pôvodných netransformovaných veličín použijeme už spomínanú mieru podobnosti, koreláciu. Pre k-means algoritmus vytvoríme obdobu matice príslušnosti. Nulovú maticu, v ktorej priradením hodnoty 1 do

j-tého stĺpca, ktorý reprezentuje j-tý zhluk, a i-tého riadku, ktorý reprezentuje i-ty prvok, budeme reprezentovať príslušnosť jednotlivých objektov do zhlukov. j-tý stĺpec bude teda veličina 'patriť do j-teho zhluku' a každý objekt bude patriť len do jedného zhluku.

Z korelačnej matice na obr.17 môžeme okrem už známych korelačných koeficientov vi-

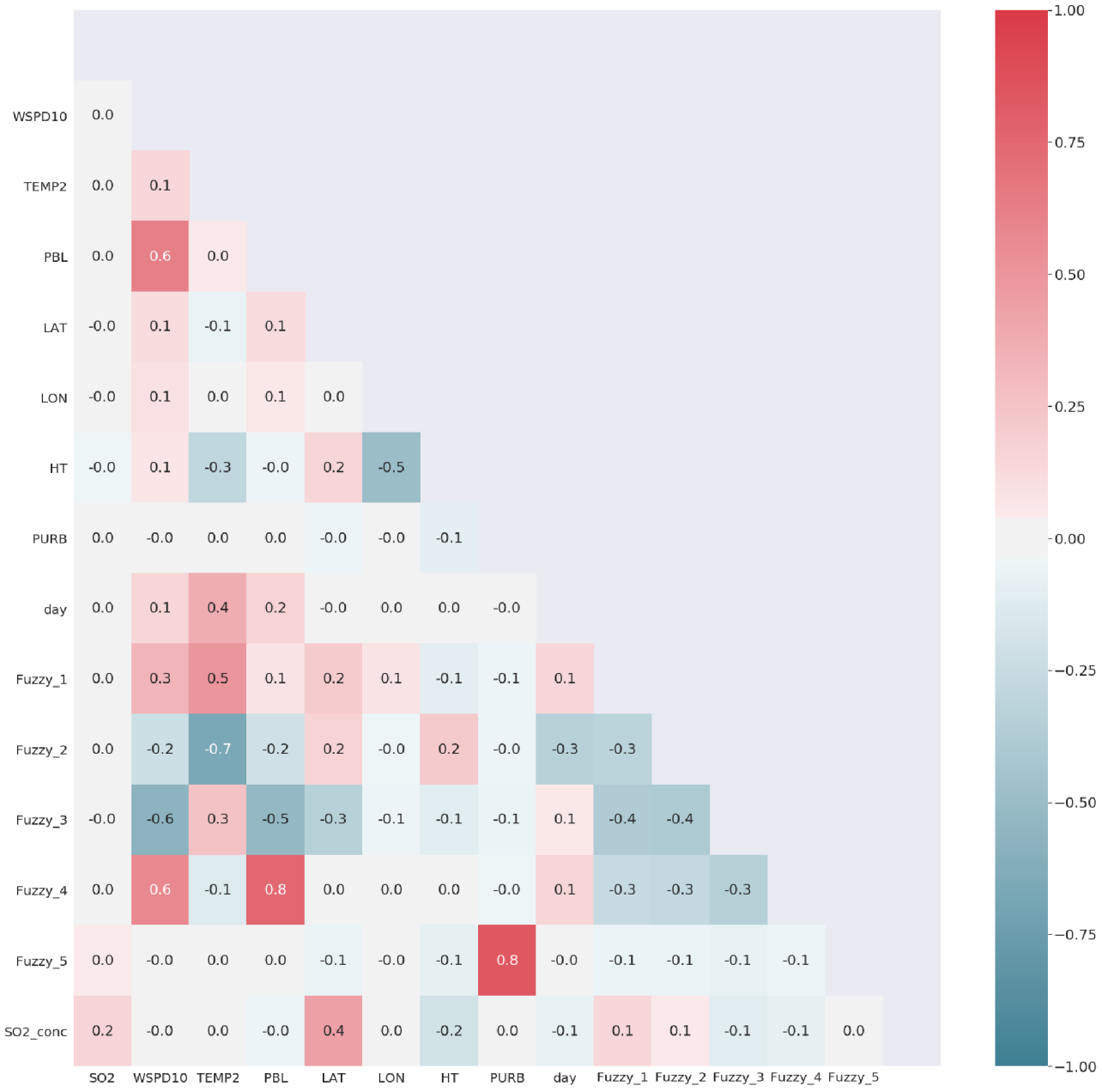


Obr. 17: Korelačná matica pre hodnoty Pearsonovho korelačného koeficientu pre dátovú maticu s pridaním matice príslušnosti k-means zhlukovania

dieť, že mnohé zhluky majú vysoké hodnoty korelačných koeficientov s veličinami, na ktorých boli zhlukované, najnižšiu najvyššiu hodnotu korelačného koeficientu má zhluk kmeans_1 s veličinou TEMP2 a to $0,5$. Najvyššie hodnoty korelačného koeficientu s pôvodným dátami majú malé zhluky. Kmeans_4 je závislí len na emisiach SO_2 a zhluk kmeans_3 je závislý na percente urbárnej zástavby v bunke. Ostatné zhluky obsahujú hodnoty korelačných koeficientov pre koreláciu s jednotlivými veličinami nižšie a väčšinou zachytávajú kombinácie rôznych vstupných veličín iných ako PURB a SO_2, preto sú tieto zhluky rovnako korelované aj vo výsledkoch pre prachové častice (v prílohe).

Obr. 18: Korelačne matica pre hodnoty Pearsonovho korelačného koeficientu pre dátovú maticu s pridaním matice príslušnosti z fuzzy c-means algoritmu s $m = 1,5$

Pre výsledky zhlukovania použitím FCM algoritmu s hodnotou váhového exponentu $m = 1.5$, do dátovej matice pridáme celú maticu príslušnosti do zhlukov, ktorá je výsledkom algoritmu, a pre tieto dáta vytvoríme korelačnú maticu na Obr.33. Z tejto matice môžme pozorovať, že závislosti medzi príslušnosťami do jednotlivých zhlukov a veličinami, ktoré sme použili na zhlukovú analýzu, sú nižšie ako pre výsledky k-means zhlukovania. Je to spôsobené väčšími zhlukmi, v ktorých nie sú veličiny, veľmi špecifikované. Zaujímavé je že zhluk k-means_3 nie je podobný s žiadnou veličinou, ktorú sme použili na jeho tvorbu. Nesie teda informáciu, ktorá je nezávislá od vstupných veličín. Ostatné zhluky majú podobné hodnoty korelačných koeficientov so vstupnými veličinami, ako zhluky ktoré vznikli metódou k-means



Obr. 19: Korelačne matica pre hodnoty Pearsonovho korelačného koeficientu pre dátovú maticu s pridaním matice príslušnosti z fuzzy c-means algoritmu s $m = 1,05$

Korelačnú maticu pre výsledky algoritmu FCM s hodnotou váhového exponentu $m = 1.05$ sme zobrazili na Obr.19. Z tejto matice môžme pozorovať, že zhluky, ktoré sú výsledkom algoritmu sa nie len na zobrazení na mape podobajú na výsledné zhluky k-means algoritmu, zhluk fuzzy_1 má rovnaké hodnoty korelačných koeficientov s veličinami z pôvodnej dátovej matice ako zhluk k-means_1,ďalej zhluk fuzzy_2 má rovnaké hodnoty koeficientov ako zhluk kmeans_1, zhluk fuzzy_3 má rovnaké hodnoty koeficientov ako zhluk k-means_6,zhluk fuzzy_4 má takmer rovnaké hodnoty koeficientov ako zhluk k-means_5 a zhluk fuzzy_5 má rovnaké hodnoty koeficientov ako zhluk kmeans_3. Jediný zhluk, ktorý nie je samostatne vyjadrený aj vo výsledkoch fuzzy c-means algoritmu je najmenší zhluk z k-means algoritmu k-means_4.

Zhrňme výsledky zhlukovej analýzy dátovej matice s veličinou PM_{10}. Výsledné zhluky, ktoré vznikli algoritmami k-means a FCM s váhovým exponentom $m = 1,5$, majú takmer také isté mohutnosti, ako zhluky, ktoré vznikli zhlukovou analýzou matice s veličinou SO_2, ako môžme vidieť v Tabuľke 9 a Tabuľke 10.

Zhluky ktoré vznikli k-means algoritmom sú takmer také isté, ako zhluky ktoré vnikli k-means algoritmom pre maticu s SO_2. Najmenší zhluk kmeans_3 má vysokú hodnotu korelačného koeficientu s veličinou PM_{10} a druhý najmenší zhluk kmeans_5 má vysokú koreláciu s percentom urbánnej zástavby, tak isto ako malé zhluky pre predošlé zhlukovanie. Ostatné zhluky sú takmer totožné, ako v predchádzajúcom zhlukovaní, čo nie je prekvapivé, keďže použitá matica bola rozdielna len pre jednu veličinu. Pre fuzzy zhlukovanie s $m = 1,5$ sú výsledné zhluky, podobné mohutnostiam aj koreláciami s veličinami, ktoré boli použité na ich vytvorenie. Žiaden zo zhlukov nemá vysokú hodnotu korelačného koeficientu s veličinou PM_{10} a zhluk fuzzy_3 tiež nemá žiadnu závislosť na zhlukovaných veličinách.

Jediný rozdiel je v rozdelení zhlukov, ktoré vznikli algoritmom FCM s váhovým exponentom $m = 1,05$, ktoré kombinujú výsledky k-means algoritmu a algoritmu fuzzy c-means s váhovým exponentom $m = 1,5$. Mohutnosti jednotlivých zhlukov sú rovnomernejšie, ale aj tak je jeden zhluk niekoľkonásobne menší, ako môžme vidieť v Tabuľke 5. Tento zhluk fuzzy_6, má vysokú koreláciu s percentom urbánnej zástavby, rovnako, ako má zhluk podobnej veľkosti k-means_5. Ostatné zhluky závisia na zhlukovaných veličinách rôzne, ale žiaden zhluk nie je úplne nezávislý. Korelačné matice, tabuľky mohutností, zobrazenie veličín na mape a zobrazenie zhlukov na mape sa nachádza v prílohe.

Tabuľka 5: Mohutnosti výsledných zhlukov algoritmu fuzzy c-means s hodnotou váhového koeficientu $m = 1,5$

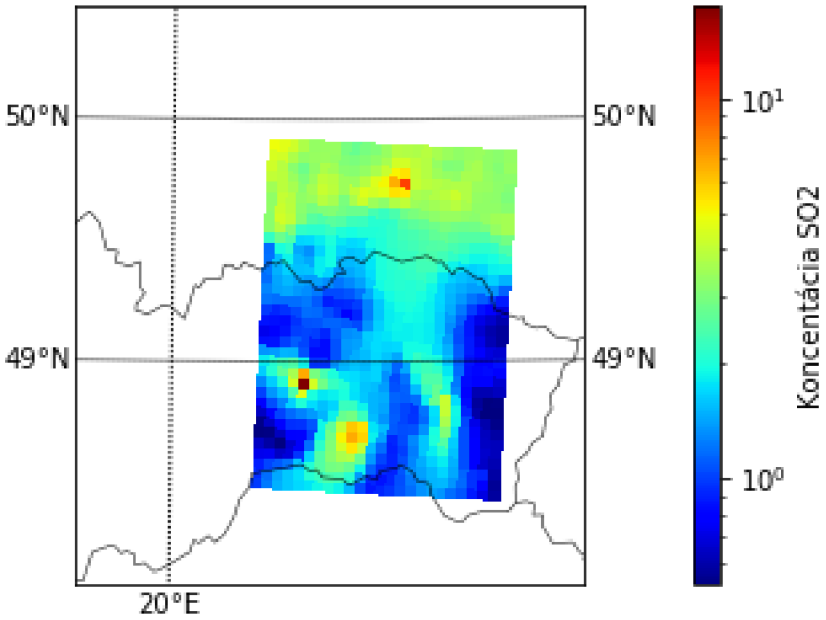
Zhluk	Mohutnosť zhluku
fuzzy_1	1676
fuzzy_2	3070
fuzzy_3	1770
fuzzy_4	976
fuzzy_5	2024
fuzzy_6	109

3.2 Regresný model

Ďalším krokom je vytvorenie štatistického modelu, ktorý bude modelovať vzťah medzi koncentráciami a emisiami. Z vizuálneho porovnania hodnôt koncentrácii a emisií oxidu siričitého na mapách na Obr. 9 a Obr. 20 môžme vidieť, že zobrazené koncentrácie na Obr. 20 majú jasné vzory v dátach. To pre emisie to neplatí, v zobrazení hodnôt na mape žiadna špecifická štruktúra, ktorá by pripomínala štruktúru v koncentráciách, nie je v dátach zjavná. Túto informáciu kvantifikujeme použitím korelačného koeficientu, ktorého hodnoty sú $r = 0,16$ a $r_s = 0,51$. Hodnota Pearsonovho korelačného koeficientu je veľmi nízka a môžeme vylúčiť lineárnu závislosť medzi emisiami a koncentráciami. Hodnota Spermanovho koeficientu je vyššia a medzi emisiami a koncentráciami je možné nájsť inú ako lineárnu závislosť. Tieto hodnoty sú však stále nízke a ak by sme na nich spravili lineárny regresný model, tak hodnota koeficientu determinácie nebude lepšia. Preto do modelu pridáme aj ďalšie veličiny, o ktorých vieme, že ovplyvňujú koncentráciu znečisťujúcich látok v atmosfére. Budú to veličiny ktoré sme použili aj na zhlukovanie. Z korelačnej matice na Obr. 17 (alebo Obr. 33), môžeme vidieť, že lineárnu závislosť $r = 0,4$ má koncentrácia oxidu siričitého so zemepisnou šírkou. Do dátovej matice pridáme

zemepisnú šírku, zemepisnú dĺžku a aj nadmorskú výšku, ktoré sú spolu s percentom urbánnej zástavby súčasťou orografických vstupov modelu.

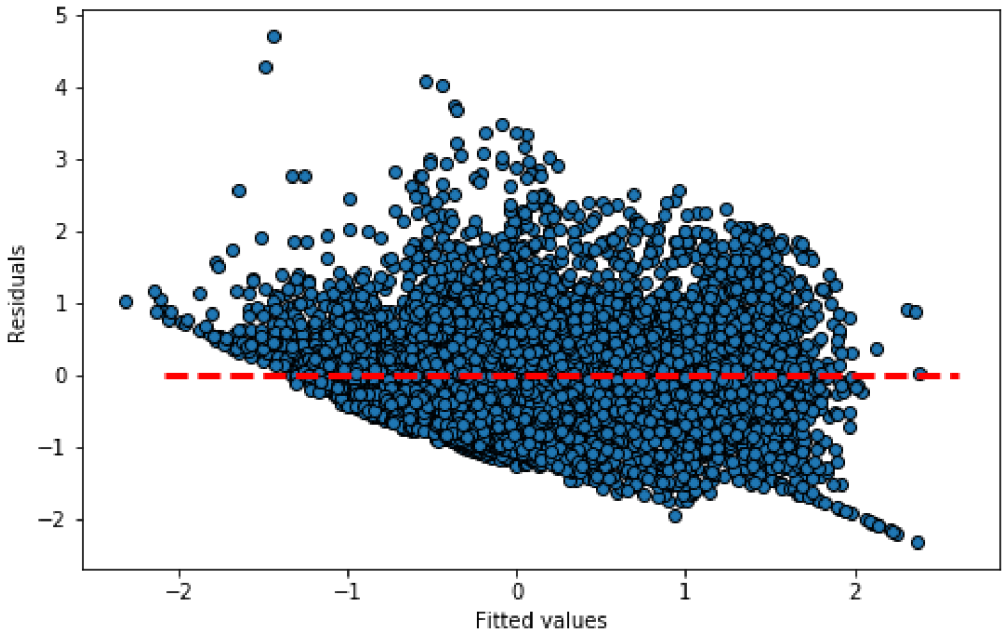
Pre takto zvolené veličiny vytvoríme kvadratický regresný model. Výsledný koeficient



Obr. 20: Hodnoty koncentrácií SO2 na zvolenej doméne pre 15. január

determinácie na ohodnotenie modelu je $r^2_{orig} = 0.529$, korigovaný koeficient determinácie $r^2_{adjorig} = 0.527$ a reziduálny súčet štvorcov $RSS = 0.686$. Graf na Obr.21 sa bežne používa na ohodnotenie nezávislostí rezíduí. Na tomto grafe môžme jasne vidieť, že tento model nespĺňa predpoklad o nezávislosti rezíduí, keďže rezíduá s rastúcim odhadom narastajú.

Máme niekoľko možností na to, aký model vytvoriť a ako použiť výsledky zhlukovej



Obr. 21: Závislosť odhadovaných hodnôt a rezíduí modelu

analýzy. Pre lineárny model na takto zvolených dátach pre všetky objekty, ktorý by zahŕňal

výsledky zhlukovania k-means algoritmu, je potrebné vyriešiť vysoké korelačné závislosti medzi zhlukovanými veličinami, ktoré nám poskytujú najmä informácie ktoré už v dátach sú, a vytvorenými zhlukmi a závislosť rezíduí. Jedno z riešení je použiť analýzu hlavných komponent, ale pre vysoké hodnoty korelačného koeficientu, ako je medzi zhlukom kmeans_2 a emisiami SO2 je vhodné jednu z premenných odobrať úplne a pre mnohé zhluky by sme úplne stratili variabilitu, ktorú nám ponúkajú. Aby sme sa vyhli vyššie uvedeným problémom a pokúsili sa vytvoriť regresný model, pre ktorý budú rezíduá nezávislé, vytvoríme regresný model pre každý zhluk samostatne.

Použitím knižnice použili knižnice [47] a [45] v programovacom jazyku python vytvoríme kvadratické regresné modely pre každý zhluk samostatne. Výsledné hodnoty koeficientov determinácie a reziduálneho súčtu štvorcov pre každý zhluk samostatne uvedieme tabuľke, v ktorej je aj hodnota výsledného koeficientu determinácie pre celý model. Pre celý model nie je dostupný korigovaný koeficient determinácie. Výsledné odhady koncentrácie vznikli spojením odhadov pre jednotlivé zhluky a nie regresným modelom.

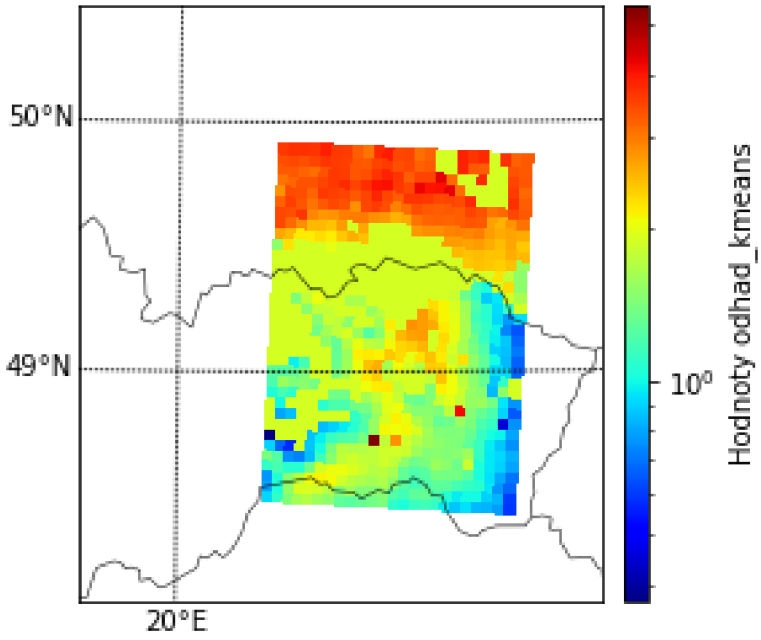
Tabuľka 6: Výsledné koeficienty determinácie kvadratického regresného modelu pre výsledné zhluky k-means algoritmu

Zhluk	r^2	r^2_{adj}	RSS
kmeans_1	0,609	0,602	0,697
kmeans_2	0,676	0,670	0,496
kmeans_3	0,725	0,551	0,448
kmeans_4	0,944	0,408	0,217
kmeans_5	0,82	0,815	0,462
kmeans_6	0,473	0,465	0,631
kmeans	0,657		0,585

V Tabuľke 6 sú výsledné hodnoty koeficientov determinácie pre regresný model na zhlukoch k-means algoritmu. Koeficienty determinácie sú pre tieto zhluky celkom vysoké, najmä pre zhluky kmeans_3 a kmeans_4, ktoré sú však najmenšie. Priemerná hodnota koeficientu determinácie je $\overline{r}^2 = 0,71$ a korigovaného koeficientu determinácie je $\overline{r}^2_{adj} = 0,59$. Takto veľký rozdiel v týchto koeficientoch je spôsobený najmä rozdielmi, ktoré sú pri koeficientoch v zhlukoch kmeans_3 a kmeans_4, pre ktoré nie je kvadratický model s toľkými závislými premennými, ako sme v ňom použili, vhodný. Ak ale vezmeme do úvahy koeficienty na väčších zhlukoch, priemerné hodnoty koeficientov sa k sebe priblížia $\overline{r}^2 = 0,64$ $\overline{r}^2_{adj} = 0,63$. Aj pre tieto priemerné hodnoty však musíme poznamenať, že hodnoty koeficientov determinácie sú lepšie pre menšie zhluky, pre najväčší zhluk nastalo zhoršenie korigovaného koeficientu determinácie.

Aby sme mohli porovnať takto vytvorený model, použijeme celkový koeficient determinácie modelu a reziduálny súčet štvorcov modelu. Vytvorený model popíše približne o $12,8\%$ viac celkovej variácie dát, ako regresný model na celej dátovej matici. Zlepšenie nastalo aj v hodnotách reziduálneho súčtu štvorcov, ktorý klesol o $= 0,101$. Aproximácie koncentrácií, ktoré sme získali takýmto modelom, sme zobrazili na mape na Obr.22.

Do Tabuľky 7 sme zapísali výsledné hodnoty koeficientov determinácie pre regresné modely na zhlukoch, ktoré vznikli fuzzy c-means algoritmom pre váhový exponent $m = 1.5$. Pre tieto zhluky sú koeficienty determinácie v porovnaní s výslednými koeficientami determinácie v Tabuľke6, na prvý pohľad nižšie. Pre žiaden zo zhlukov koeficient determinácie nepresahuje hodnotu 0.85, a že pre najväčší zhluk fuzzy_5 hodnota korigovaného koeficientu determinácie, podobne ako pre kmeans_6, klesla. Taktiež si všimneme, že najme vďaka rovnomernejšie roz-



Obr. 22: Hodnoty odhadu koncentrácií SO2 na zvolenej doméne pre 15. január, získane regresným modelom pre zhluky metódy k-means

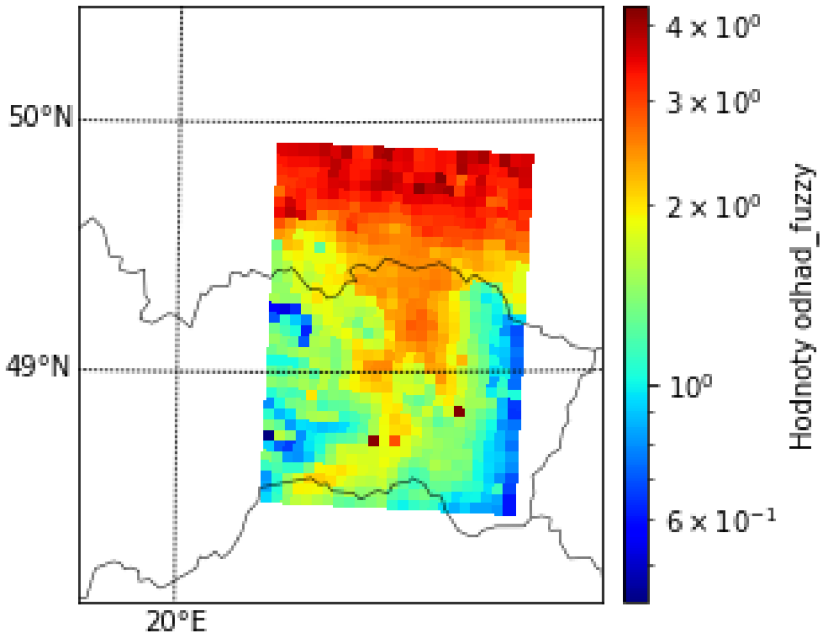
deleným zhlukom sú menšie rozdieli medzi r^2 a r^2_{adj} .

Pri porovnaní hodnôt koeficientu determinácie, modelu, ktorý vznikol spojením odhadov regresných modelov pre samostatné zhluky s pôvodnou hodnotou zistíme, že model, ktorý vznikol spojením odhadov, popíše približne o $13,3\%$ viac z celkovej variability dát. Podobné zlepšenie vidíme aj v hodnotách reziduálneho súčtu štvorcov, ktorý klesol o $0,106$. Aproximácie koncentrácií, ktoré sme získali takýmto modelom, sme zobrazili na mape na Obr.23

Tabuľka 7: Výsledné koeficienty determinácie kvadratického regresného modelu pre výsledné zhluky fuzzy c-means algoritmu pre váhový exponent $m = 1.5$

Zhluk	r^2	r^2_{adj}	RSS
fuzzy_1	0,691	0,682	0,515
fuzzy_2	0,833	0,824	0,414
fuzzy_3	0,695	0,687	0,486
fuzzy_4	0,562	0,550	0,760
fuzzy_5	0,418	0,408	0,656
fuzzy_6	0,826	0,820	0,425
$\text{fuzzy}_{m=1,5}$	0,662		0,580

Výsledné hodnoty koeficientov determinácie pre regresný model na zhlukoch, ktoré vznikli fuzzy c-means algoritmom pre váhový exponent $m = 1,05$, sú uvedené v Tabuľke 8. Pre tieto zhluky sú koeficienty determinácie nižšie ako pre zhluky, ktoré vznikli metódou k-means, aj napriek tomu, že sa im podobajú. Hodnotami koeficientov determinácie sa blížia koeficientom determinácie v Tabuľke 7, ktorá obsahuje súčet štvorcov rezíduí a koeficienty determinácie pre



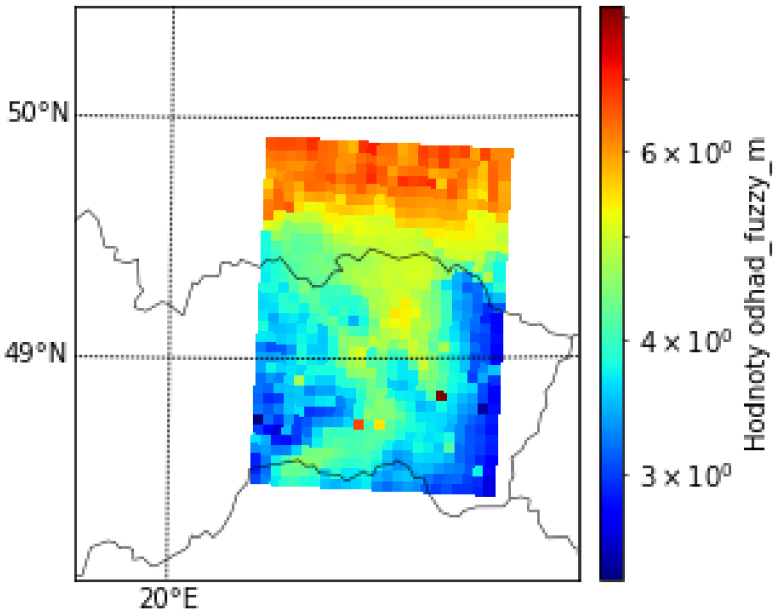
Obr. 23: Hodnoty odhadu koncentrácií SO2 na zvolenej doméne pre 15. január, získane regresným modelom pre zhluky metódy fuzzy c-means

podmodely, ktoré vznikli na jednotlivých zhlukoch (a aj pre celkový model) FCM s váhovým exponentom $m = 1.5$. Pre žiaden zo zhlukov koeficient determinácie nepresahuje hodnotu $0,80$ a medzi hodnotami korigovaných koeficientov determinácie a koeficientami determinácie je pre toto delenie menší rozdiel, priemerná hodnota koeficientu determinácie je $\overline{r}^2 = 0,65$ a hodnota korigovaného koeficientu determinácie je $\overline{r}^2_{adj} = 0,61$. Aproximácie koncentrácií, ktoré sme získali takýmto modelom, sme zobrazili na mape na Obr.23

Tabuľka 8: Výsledné koeficienty determinácie kvadratického regresného modelu pre výsledné zhluky fuzzy c-means algoritmu pre váhový exponent $m = 1.05$

Zhluk	r^2	r^2_{adj}	RSS
fuzzy_1	0,599	0,591	0,705
fuzzy_2	0,671	0,664	0,499
fuzzy_3	0,474	0,466	0,633
fuzzy_4	0,788	0,783	0,499
fuzzy_5	0,725	0,551	0,448
$\text{fuzzy}_{m=1,05}$	0,645		0,596

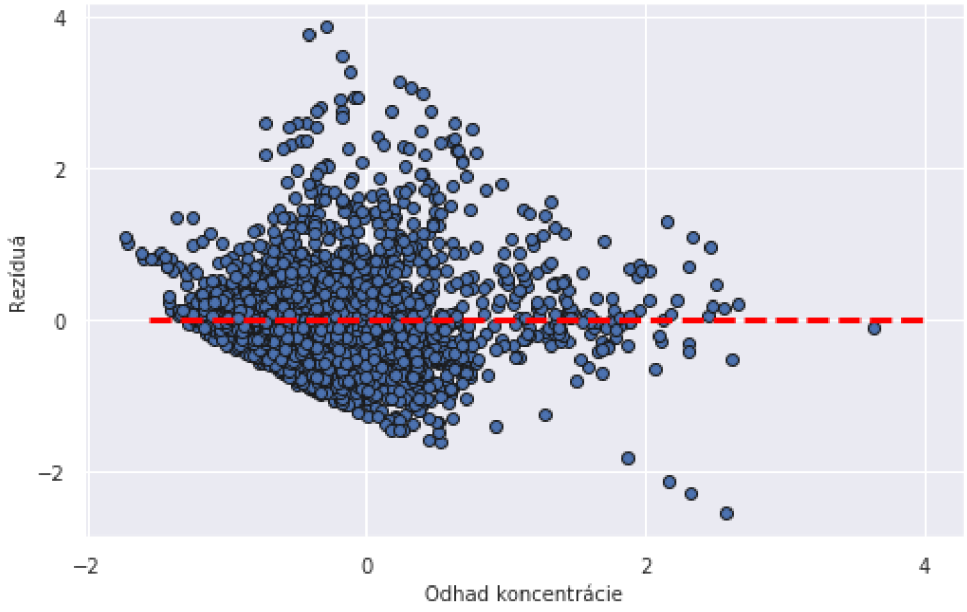
To čo nás zaujíma najviac je porovnanie medzi modelmi, ktoré vznikli použitím zhlukov z rôznych algoritmov. Keď porovnáme tieto modely medzi sebou, môžme viedieť, že hodnoty koeficientov determinácie a reziduálneho súčtu štvorcov, sú si veľmi blízke. Najhoršiu hodnotu r^2 má model, ktorý vznikol na zhlukoch, ktoré vznikli algoritmom fuzzy c-means pre váhový exponent $m = 1,05$, čo je prekvapivé. Je to model, pre ktorý sme práve mohli očakávať, že keďže je to kombinácia výsledku k-means algoritmu a fuzzy c-means algoritmu, výsledky sa na ňom zlepšia. Pre tento model sme ale zmenili aj počet zhlukov, na základe zvýšenia hodnoty



Obr. 24: Hodnoty odhadu koncentrácií SO2 na zvolenej doméne pre 15. január, získane regresným modelom pre zhluky metódy fuzzy s-means s $m = 1,05$

fpc na Obr.13.Najvyššiu presnosť modelu sme získali práve na na zhlukoch, ktoré vznikli algoritmom fuzzy c-means pre váhový exponent $m = 1,5$, ktoré boli najrovnomernejšie rozdelené.

Keď sa pozrieme na grafy rezíduí pre jednotlivé zhluky, ktoré vznikli algoritmom k-means

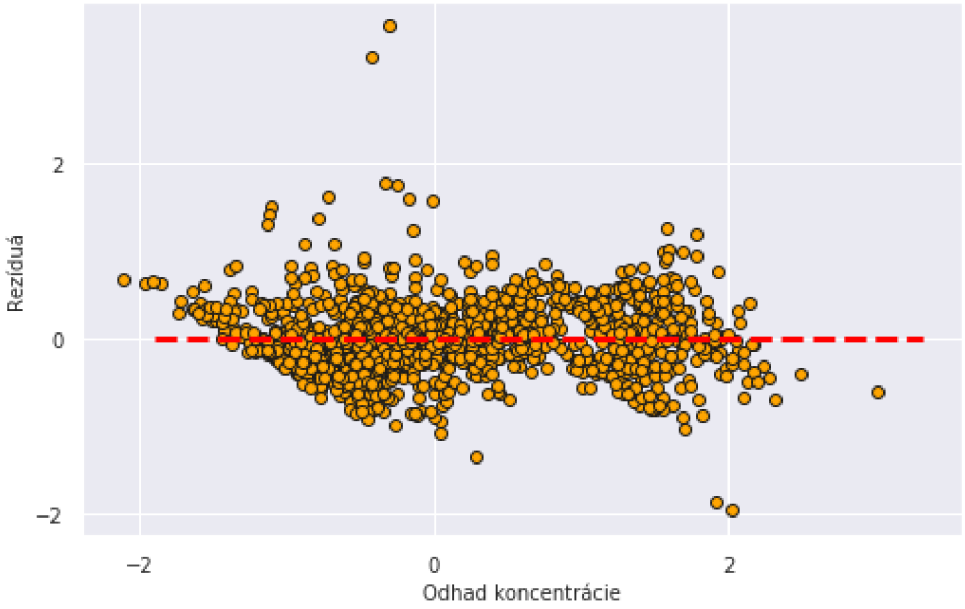


Obr. 25: Graf závislostí rezíduí na odhade koncentrácie SO_2 kvadratického regresného modelu na zhluku kmeans_6

alebo mali hodnotu váhoveho exponentu $m = 1,05$, je na všetkých grafoch viditeľná závislosť rezíduí na odhade koncentrácií SO_2, ako je to napr. na Obr.25. Pre zhluky, ktoré vzniky algoritmom FCM s hodnotou váhového exponentu podla Bezdeka, je pre väčšinu zhlukov na reziduálnom grafe prítomná podobná závislosť. Reziduálne grafy pre dva zhluky fuzzy_3 a

fuzzy_6 na Obr.26, tvary, ktoré by naznačovali závislosť medzi rezíduami nižšia. Zaujímavé je to, že na zhluku fuzzy_3, ktorý nezávisel na žiadnej vstupnej veličine, sú nezávislé reziduá regresného modelu a veľmi nízka hodnota koeficientu determinácie. Dokonca nižšia ako mal pôvodný model.

Vyhodnotiť vhodnosť modelov nie je jednoduché, najme kvôli marginálnym rozdielom



Obr. 26: Graf závislostí rezíduí na odhade koncentrácie SO_2 kvadratického regresného modelu na zhluku fuzzy_6

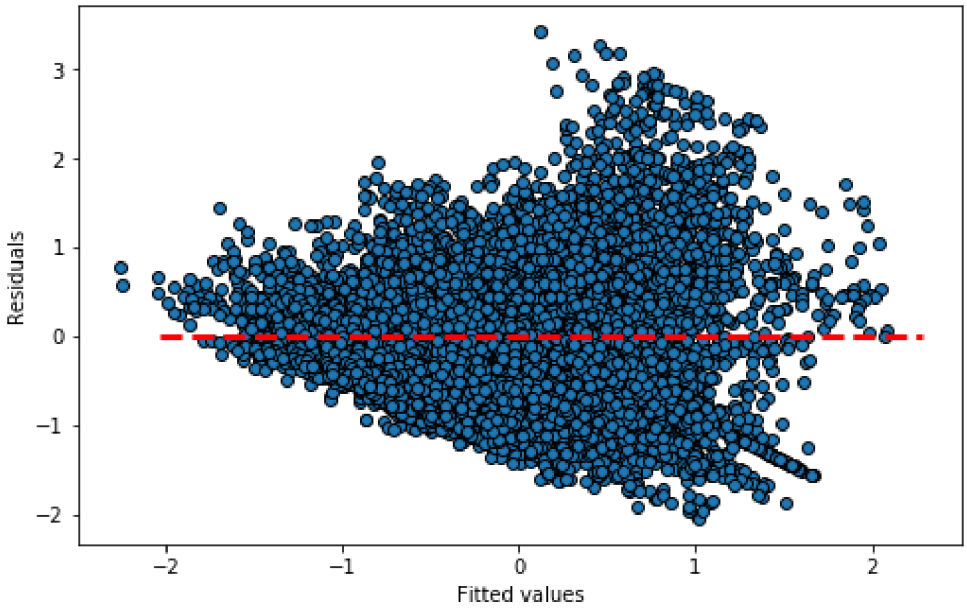
v hodnoteniach modelu. Reziduálne grafy sú lepšie pre zhluky z FCM algoritmu s hodnotou váhového exponentu $m = 1,5$. Vo výsledku je teda takto vytvorený model najlepší z pomedzi vytvorených modelov.

Vyhodnoťme aj použitie regresnej analýzy na nájdenie vzťahu medzi emisiami a koncentráciami PM_{10}. Korelácie medzi emisiami a koncentráciami znečisťujúcej látky PM_{10} je nižšia ako pre SO_2. Korelačné koeficienty majú hodnotu len $r_s = 0,16$ a Pearsnov $r = 0,05$,a teda v dátach nie je žiadna očividná závislosť.

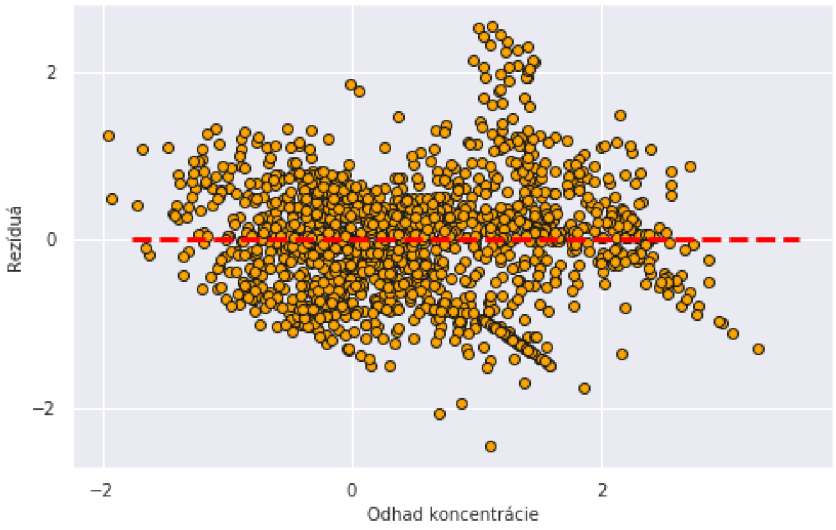
Regresnú analýzu aplikujeme na dátovú maticu pre PM_{10}, ktorá je taká istá ako sme použili pre regresnú analýzu pre prvok SO_2. Táto dátová matica nezávislých premenných regresného modelu bude tvorená veličinami PM_{10}, WSPD10 (rýchlosť vetra), TEMP2 (teplota vo výške 2m), PBL (výška premiešavania), PURB (podiel mestskej zástavby), LAT (zemepisné šírka), LON (zemepisné dĺžka) a HT (nadmorská výška). Závislú premennú bude predstavovať hmotnostná koncentrácia prachových častíc, táto premenná sa v dátovej matici nazýva PM10_conc. Pre tieto veličiny je výsledný koeficient determinácie $r^2 = 0.423$ a korigovaný koeficient determinácie $r^2 = 0.421$. Tieto hodnoty sú výrazne nižšie, ako pre regresnú analýzy s koncentráciou SO_2, keďže vzťah medzi koncentráciami a emisiami je komplikovanejší a emisie PM_{10} zostávajú v doméne dlhšie. Závislosť reziduí na odhade hodnoty PM10_conc nie je splnená ani pre tento model. Vhodnejší by bol iný komplexnejší štatistický model, alebo viac menších modelov. Tiež sme preto pristúpili k regresným modelom na zhlukoch.

Tabuľky s podrobnejšími výsledkami sa nachádzajú v prílohe. Môžeme v nich pozorovať, že podobne ako v tabuľkách pre SO_2, výsledné koeficienty determinácie sú vyššie pre zhluky s malou mohutnosťou, ale korigované koeficienty sú pre ne veľmi nízke a pre zhluk kmeans_3 dokonca nie je možné korigovaný koeficient determinácie vypočítať vôbec.

Výsledný koeficient determinácie aplikácie rôznych regresných modelov na zhluky s pev-



Obr. 27: Graf závislostí rezíduí na odhade koncentrácie PM_{10} kvadratického regresného modelu na pôvodných dátach



Obr. 28: Graf závislostí rezíduí na odhade koncentrácie PM_{10} kvadratického regresného modelu pre zhluk fuzzy_4

nou štruktúrou, ktoré sú výsledkom k-means algoritmu je $r^2 = 0,585$, pre zhluky trochu fuzzy štruktúrou, ktoré sú výsledkom FCM s váhovým exponentom $m = 1,05$ je $r^2 = 0,592$ a nakoniec pre zhluky, kotých zhluky predstavujú skutočne fuzzy množiny je to $r^2 = 0,613$. V tomto prípade môžme vidieť, že zo stúpajúcim váhovým exponentom sa zväčšuje aj koeficient determinácie. Čím bližšie sú mohutnosti zhlukov k rovnomernému rozdeleniu, tým vyššie sú hodnoty koeficientu determinácie, čo je samozrejme prirodzené, lebo je jednoduchšie vytvoriť jednoduchý model, ktorý zachytáva vzťah medzi nezávislou a závislou premennou na nižšom počte objektov. Modely vytvorené na získanie odhadu koncentrácie PM_{10} majú lepšie výsledky (Obr.28) nezávislosti rezíduí ako modely, ktoré sme vytvorili pre SO_2.

Keď navzájom porovnáme zvýšenie presnosti odhadu koncentrácii medzi SO_2 a PM_{10}, vy-

tvorenie zhlukov a samostatných regresných modelov malo pre PM_{10} väčší efekt. Konkrétny spôsob rozdelenia dát do zhlukov, však mal len malý efekt na výsledný koeficient determinácie pre obe odhadované veličiny.

3.3 Možnosti ďalšieho vývoja práce

Zvýšenie váhového exponentu

Z výsledkov regresnej analýzy môžme usúdiť, že koeficient determinácie sa zlepšoval s väčším počtom rovnomerne mohutných zhlukov. Preto je vhodné uvažovať o ďalšom zvýšení váhového exponentu, čo by viedlo k ešte rovnomernejšie rozdeleným zhlukom, aj keď by to znamenalo stratu štruktúry zhlukov.

Aplikácia na celoročné dáta

Prirodzením pokračovaním práce je rozšírenie modelu na pôvodnú doménu a rozšírenie časového intervalu. Toto je však náročnejšia úloha, ktorá bude vyžadovať aj značné rozšírenie použitých premenných, aby sme zachytili vzťah medzi koncentráciami a premennými. V takom prípade už bude nutné použiť hodnoty smeru vetra a mokrej depozície čo znemožňuje použitie denných priemerov a bude potrebné použiť hodinové údaje. Podstatný je tiež vplyv atmosférických zrážok.

Použitie komplexnejších modelov

V tejto práci sme sa venovali porovnaniu získaniu rôznych zhlukov a nájdeniu vhodného modelu na odhad koncentrácii sme nevenovali veľkú pozornosť. Je to však jedná z častí na ktorej je potrebné ďalej pracovať, keďže mnohé z vytvorených regresných modelov nespĺňajú predpoklad o normalite rezíduí. Okrem možnosti využitia komplexnejších predikčných modelov ako sú zložitejšie regresné modely, FIS, neurónove siete a iné, je tiež možnosť na každý model nájsť vhodný model samostatne.

Nepoužiť zhlukovú analýzu

Vo výsledku ponúkla zhluková analýza len malé zlepšenie v presnosti odhadu koeficientu determinácie regresných modelov. Najvýraznejšie sa tieto hodnoty zlepšili pre rovnomerne rozdelené zhluky, čo nie je prekvapivé. Je len otázne, či by sa táto hodnota naďalej zlepšovala s rovnomernosťou delenia, alebo či by pre náhodne delenie klesla. Výpočet príslušnosti do rôznych zhlukov je tiež výpočtovo náročný, a aj napriek tomu, že rozdelenie do zhlukov zlepšilo popísanie variácie dát PM_{10} modelom až o 19%, možno by bolo vhodnejšie aplikovať na celú doménu sofistikovanejší model, keďže ani modely na zhlukoch nespĺňali základné predpoklady regresnej analýzy.

Záver

V prvej kapitole sme sa venovali základným poznatkom o zhlukovej analýze. Uviedli sme niekoľko algoritmov, ktoré sa používajú v praxi. Špeciálne sme sa venovali nehierarchickým metódam a segmentačným metódam založeným na neurónových sieťach, ktoré sú založené na rozšírení klasickej metódy o fuzzy množiny alebo fuzzy logiku. Tieto metódy sú fuzzy c-means, ktorý je rozšírením zhlukovacej metódy k-means, a fuzzy ART, ktorý je rozšírením metódy súťaživého učenia ART. V závere kapitoly sme zhrnuli teóriu o regresných modeloch, ktoré sme použili v závere tretej kapitoly tejto diplomovej práce.

V druhej kapitole sme zhrnuli základné poznatky o modelovaní kvality ovzdušia. Popísali sme niektoré chemické procesy, ktorým podliehajú znečisťujúce látky v atmosfére, a popísali sme rôzne druhy druhy modelovania koncentrácii týchto látok podľa použitého rozlíšenia. Veľká časť druhej kapitoly sa venuje niekoľkým typom modelov, ktoré sa používajú na výpočet (odhad) koncentrácii znečisťujúcich látok v atmosfére. Detailnejšie sme popísali chemicko transportný model CMAQ a jeho moduly. Vstupné a výstupné dáta tohto modelu sme ďalej analyzovali v tretej kapitole.

V poslednej tretej kapitole sme previedli praktickú časť. Zhlukovú analýzu sme aplikovali na dve rôzne dátové matice. Prvá, ktorej sa venujeme v kapitole detailnejšie, mala ako jednu z veličín (premenných modelu) oxid siričitý SO_2 , druhá mala ako jednu z veličín prachové častice PM_{10}, ostatné veličiny boli pre matice totožné (PBL, teplota, rýchlosť vetra, podiel mestskej zástavby). Tieto matice a jednotlivé veličiny sme po písali a spravili sme na nich základnú štatistická analýzu. Tieto dáta sme pred spracovali (pre-processing) nutnými transformáciami, aby na nich bolo možné spraviť zhlukovú analýzu a heuristickými metódami sme na základe dátovej štruktúry zvolili optimálny počet zhlukov pre k-means zhlukovanie a fuzzy c-means (FCM) zhlukovanie. Vizualizácia fuzzy partition koeficietu (FPC) 13 nemá skok, ktorý by pre nejakú hodnotu váhového exponentu m jednoznačne určil optimálny počet zhlukov, preto sme porovnali dva rôzne prístupy, ktoré sú odporúčané na zvolenie minimálnej hodnoty m. Výsledky zhlukovacích metód sme vyhodnotili. Metóda k-means pre obidve matice našla jasnú ale nerovnomernú štruktúru. Výsledky FCM pre menšiu hodnotu váhového exponentu $m = 1,05$ mali veľmi podobnú štruktúru ako zhluky, ktoré boli vytvorené metódou k-means, pre vyššiu hodnotu váhového exponentu $m = 1,5$ boli výsledné zhluky menej štrukturované a v niektorých prípadoch dokonca nemali žiadnu podobnosť zo zhlukovanými veličinami.
Na konci tretej kapitoly sme na zhlukoch, ktoré sme vytvorili zhlukovou analýzou vytvorili regresné modely, ktoré sme porovnali navzájom a s regresným modelom na celej matici. Model, ktorý sme spravili spojením regresných modelov, ktoré sme spravili na jednotlivých zhlukoch, ma lepšie výsledky ako model na celej matici, ale nie oveľa. Celkovo najlepší týmto spôsobom vznikol na zhlukoch ktoré boli výsledkom FCM metódy s $m = 1,5$. Tieto modely poskytli vylepšenie, ale nie významné a na zhlukoch implementované modely stále nespĺňali predpoklady nezávislosti rezíduí na odhadovanej hodnote. Na tieto dáta je teda nutné aplikovať komplexnejší model (napríklad neurónovú sieť), ktorý by dokázal lepšie zachytávať vzťah medzi koncentráciami a emisiami. V takom prípade bude zhluková analýza mať iný efekt.

Literatúra

[1] ANDĚL, Jiří. Statistické metody. Matfyzpress, 2007.ISBN 80-7378-003-8

[2] ANDĚL, Jiří. Základy matematické statistiky. Univerzita Karlova v Praze, Matematicko-fyzikální fakulta, 2002.ISBN 80-2701-712-2

[3] BEZDEK, James C.; EHRLICH, Robert; FULL, William. FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 1984, 10.2-3: 191-203.

[4] BEZDEK, James C.,HATHAWAY, Richard J.,SABIN, Michael J., and William T. TUCKER. "Convergence theory for fuzzy c-means: counterexamples and repairs.ÏEEE Transactions on Systems, Man, and Cybernetics 17, no. 5, 1987 : 873-877.

[5] BYUN, Daewon; SCHERE, Kenneth L. Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system. 2006.

[6] CABANEROS, Sheen Mclean; John Kaiser; HUGHES, Ben Richard. A review of artificial neural network models for ambient air pollution prediction, Environmental Modelling & Software, 285-304. ISSN 1364-8152, https://doi.org/10.1016/j.envsoft.2019.06.014.

[7] CANNON, Robert L.; DAVE, Jitendra V.; BEZDEK, James C. Efficient implementation of the fuzzy c-means clustering algorithms. IEEE transactions on pattern analysis and machine intelligence, 1986, 2: 248-255.

[8] DAS, Samarjit; BARUAH, Hemanta K. Application of fuzzy c-means clustering technique in vehicular pollution. Journal of Process Management. New Technologies, 2013, 1.3: 94-105.

[9] DOGRUPARMAK, Senay Cetin, et al. Using principal component analysis and fuzzy c–means clustering for the assessment of air quality monitoring. Atmospheric Pollution Research, 2014, 5.4: 656-663.

[10] DRAPER, N.R. and SMITH, H. Applied regression analysis 3.rd edition. John Wiley & Sons. 1998. ISBN 0-471-17082-8

[11] DU, K-L. "Clustering: A neural network approach."Neural networks 23, no. 1, 89-107. 2010.

[12] DUNEA, Daniel; POHOATA, Alexandru Alin; LUNGU, Emil. Fuzzy inference systems for estimation of air quality index. ROMAI Journal, Romanian Society of Applied and Industrial Mathematics, 2011, 7.2: 63-70.

[13] ENVIRON. User's Guide to the Comprehensive Air Quality Model with Extensions (CAMx) Version 2.00. ENVIRON. 1998. International Corporation, 101 Rowland Way, Suite 220, Novato, California 94945-5010

[14] GAN, Guojun; MA, Chaoqun; WU, Jianhong. Data clustering: theory, algorithms, and applications. Siam, 2007.ISBN 978-0-898716-23-8

[15] GOYAL, P.; KUMAR, Anikender. Mathematical modeling of air pollutants: an application to Indian urban city. Air Quality-Models and Applications, 2011, Prof. Dragana Popovic (Ed.), ISBN: 978-953-307-307-1, InTech,[online] Dostupné z: http://www.intechopen.com/books/air-quality-models-and-applications/mathematical-modeling of-air-pollutants-an-application-to-indian-urban-city

[16] GROVE, Robert, Using the elbow method to determine the optimal number of clusters for k-means clustering citovane[online]. 2017 [cit 2020-6-17]. Dostupné z https://bl.ocks.org/rpgove/0060ff3b656618e9136b

[17] HAMEDIAN, Amir Abbas, et al. Air Quality Analysis by Using Fuzzy Inference System and Fuzzy C-mean Clustering in Tehran, Iran from 2009–2013. Iranian journal of public health, 2016, 45.7: 917.

[18] HING, Jason; BYUN, Daewon. Introduction to the Models-3 framework and the Community Multiscale Air Quality model (CMAQ). Science Algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) Modeling System, 1999.

[19] IZENMAN, Alan Julian. Modern multivariate statistical techniques. Regression, classification and manifold learning, 2008. ISBN 978-0-387-78188-4

[20] KINGSY, Grace R., et al. Air pollution analysis using enhanced K-Means clustering algorithm for real time sensor data. In: 2016 IEEE Region 10 Conference (TENCON). IEEE, 2016. p. 1945-1949.

[21] KOLESÁROVÁ, Anna; KOVÁČOVÁ, Monika. Fuzzy množiny a ich aplikácie. Slovenská technická univerzita, 2004. ISBN 80-227-2036-4

[22] LUKASOVÁ, Alena; ŠARMANOVÁ, Jana. Metody shlukové analýzy. Státní nakladatelství technické literatury, 1985.

[23] MAILLER, Sylvain, et al. CHIMERE-2017: from urban to hemispheric chemistry-transport modeling. 2017. Geosci. Model Dev., 10, 2397-2423.Dostupné z: https://doi.org/10.5194/gmd-10-2397-2017.

[24] NAVARA, Mirko; OLŠÁK, Petr. Základy fuzzy množin. České vysoké učení technické, 2002. ISBN 80-01-02585-3

[25] NOONE, Kevin. Atmospheric Chemistry and Physics: From Air Pollution to Climate Change, by John H. Seinfeld and Spyros N. Pandis. Physics Today, 1998, 51: 88-90 [online]. Dostupné z: https://doi.org/10.1063/1.882420

[26] NOVÁK, Vilém. Základy fuzzy modelování. Praha: BEN-technická literatura. 2000. ISBN 80-7300-009-1.

[27] OKE, Timothy R., et al. Urban climates. Cambridge University Press, 2017 294-331 [online] dostupné z: https://doi.org/10.1017/9781139016476.012

[28] PAL, Nikhil R.; BEZDEK, James C. On cluster validity for the fuzzy c-means model. IEEE Transactions on Fuzzy systems, 1995, 3.3: 370-379.

[29] RESLER, Jaroslav; KRČ, Pavel; BELDA, Michal; JURUS, Pavel; BENEŠOVÁ, Nina; LOPATA, Jan; VlČEK, Ondřej; DAMAŠKOVÁ, Daša; EBEN, Krystof; DERBEK, Přemysl; MARONGA, Björn; KANANI-SUHRING, Farah. A new urban surface model integrated in the large-eddy simulation model PALM. Geoscientific Model Development Discussions. 1-26. 10.5194/gmd-2017-61. 2017.

[30] Správa o kvalite ovzdušia v Slovenskej republike 2019, vydal SHMÚ, 111 pp., v tlači

[31] SILVEIRA, Carlos, FERREIRA, Joana a Miranda, ANA. (2019). The challenges of air quality modelling when crossing multiple spatial scales. Air Quality, Atmosphere & Health. 12. 10.1007/s11869-019-00733-5.

[32] STEIN, A. F., et al. NOAA's HYSPLIT atmospheric transport and dispersion modeling system. Bulletin of the American Meteorological Society, 2015, 96.12: 2059-2077. Dostupné z http://dx.doi.org/10.1175/BAMS-D-14-00110.1

[33] TRAUWAERT, E. Ön the meaning of Dunn's partition coefficient for fuzzy clusters."Fuzzy sets and systems 25, no. 2 1988: 217-242.

[34] TRAVNIKOV, Oleg; ILYIN, Ilia. The EMEP/MSC-E mercury modeling system. In: Mercury Fate and Transport in the Global Atmosphere. Springer, Boston, MA, 2009. p. 571-587.

[35] XU, Rui; WUNSCH, Don. Clustering IEEE Press Series on Computational Intelligence – Volume 10. John Wiley & Sons, 2009. ISBN 978-0-470-27680-8

[36] YU, Jian; CHENG, Qiansheng; HUANG, Houkuan. Analysis of the weighting exponent in the FCM. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2004, 34.1: 634-639.

[37] ZADEH, Lotfi A. Fuzzy sets. Information and control, 1965, 8.3: 338-353.

[38] ZÁVODSKÝ, D.; MEDVEĎ, M.; ĎUREC, F. Chémia atmosféry a modelovanie znečisťovania ovzdušia. Univerzita Mateja Bela, Banská Bystrica, 2001 ISBN 80-88784-34-4.

[39] ZVÁRA, Karel. Regrese Matfyzpress, 2007. ISBN 978-80-7378-041-8

[40] SCIKIT-FUZZY manuál [online].[cit. 2020-06-22]. Dostupné z https://github.com/scikit-fuzzy/scikit-fuzzy

[41] MATPLOTLIB manuál [online].[cit. 2020-06-22]. Dostupné z: https://matplotlib.org/contents.html

[42] NUMPY manuál [online].[cit. 2020-06-22]. Dostupné z: https://docs.scipy.org/doc/numpy/dev/

[43] PANDAS manuál [online].[cit. 2020-06-22]. Dostupné z: https://pandas.pydata.org/pandas-docs/stable/

[44] SEABORN manuál [online].[cit. 2020-06-22]. Dostupné z: https://seaborn.pydata.org/tutorial.html#tutorial

[45] SKLEARN manuál [online].[cit. 2020-06-22]. Dostupné z: https://scikit-learn.org/stable/documentation.html

[46] SCIPY manuál [online].[cit. 2020-06-22]. https://www.scipy.org/scipylib/index.html

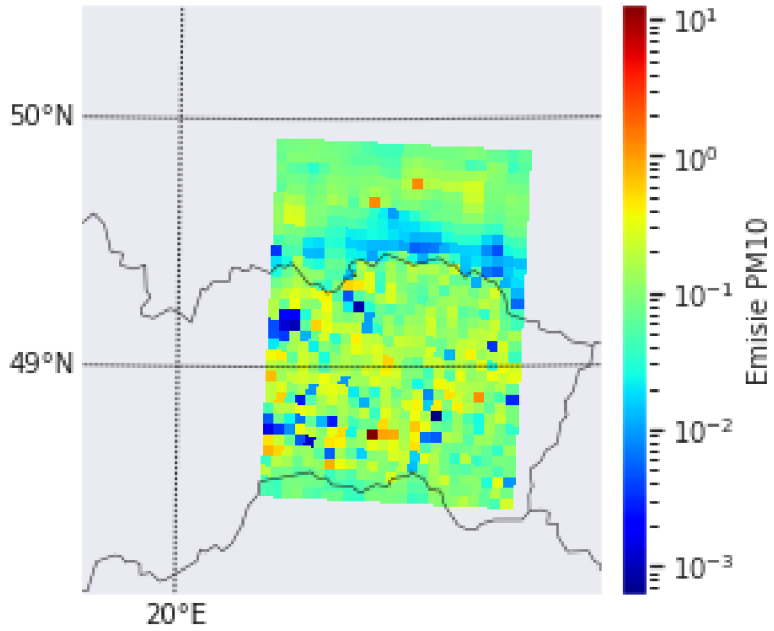
[47] STATSMODELS manuál [online].[cit. 2020-06-22]. Dostupné z: https://www.statsmodels.org/stable/index.html

ZOZNAM PRÍLOH

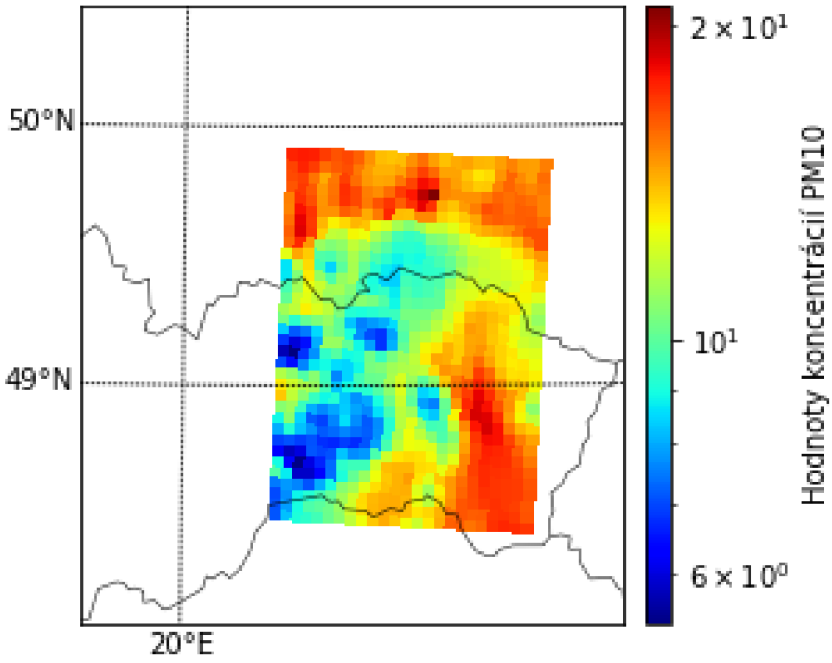
A.1: Výsledky a zobrazenia pre PM_{10}
B.1: Dáta[1]
B.2: Výsledky zhlukovej analýzy, vizualiácie veličín a zhlukov na mape[2]
B.3: Výsledku regresného modelu, vizualizácie koncentrácii na mape[3]

[1] V priečinku DP_Camara
[2] V priečinku DP_Camara
[3] V priečinku DP_Camara

A.1: Výsledky a zobrazenia pre PM_{10}



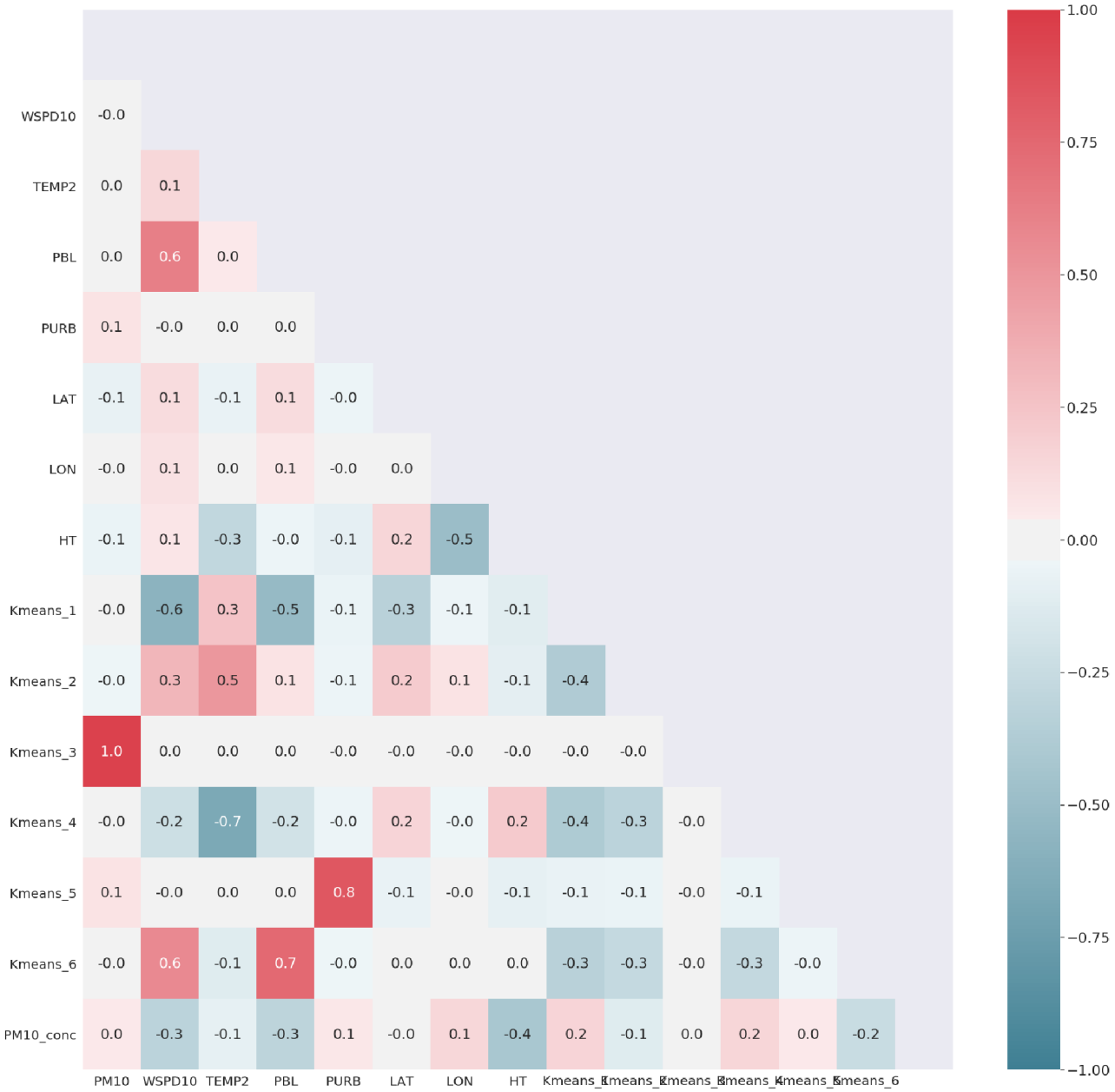
Obr. 29: Hodnoty emisií PM_{10} na zvolenej subdoméne pre 15. január. Vstup modelu CMAQ.



Obr. 30: Hodnoty koncentrácií PM_{10} na zvolenej subdoméne pre 15. január. Výstup modelu CMAQ.

Tabuľka 9: Mohutnosti výsledných zhlukov algoritmu k-means s použitím veličiny PM_{10}

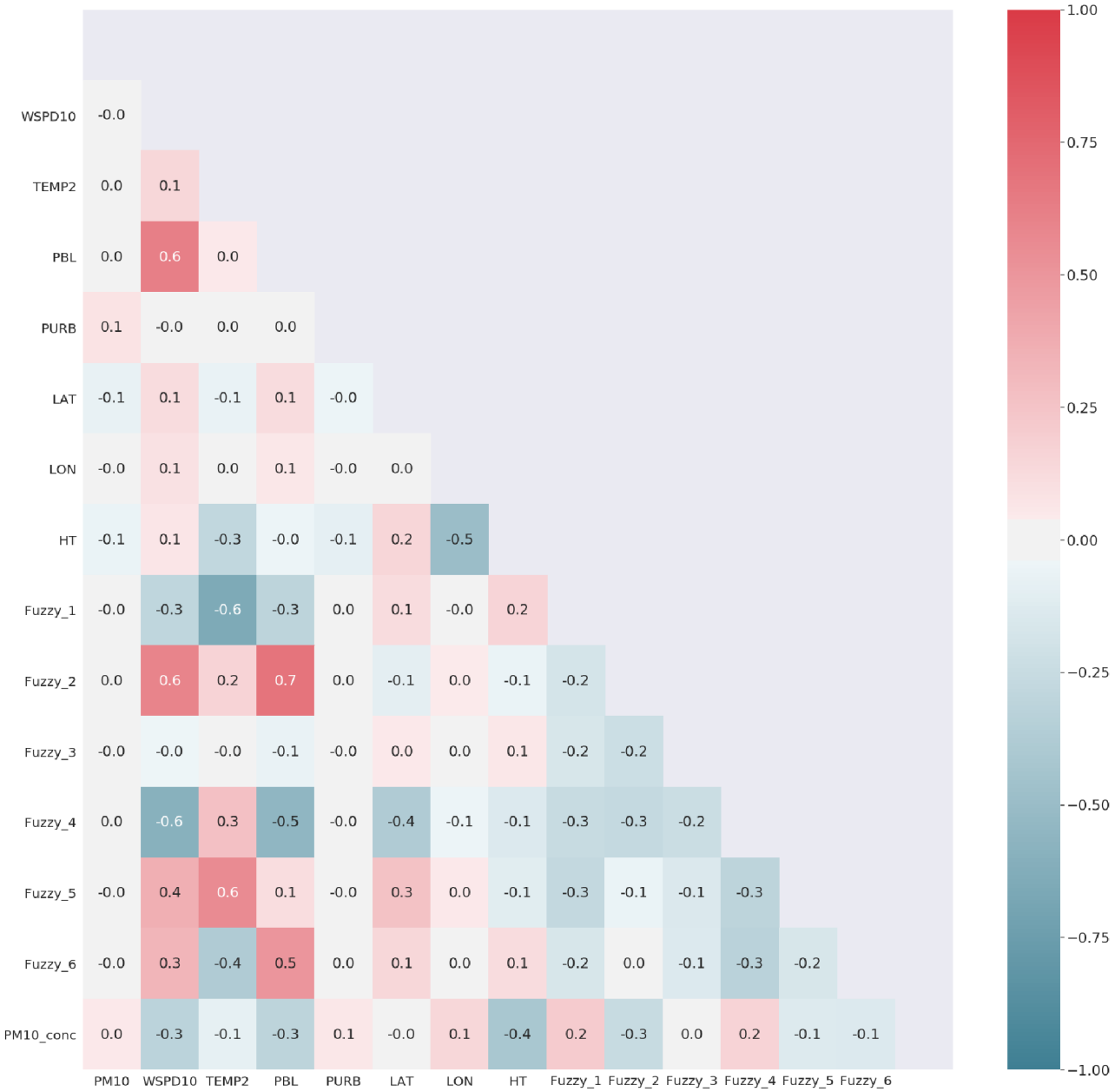
Zhluk	Mohutnosť zhluku
kmeans_1	3081
kmeans_2	2254
kmeans_3	11
kmeans_4	2332
kmeans_5	99
kmeans_6	1848



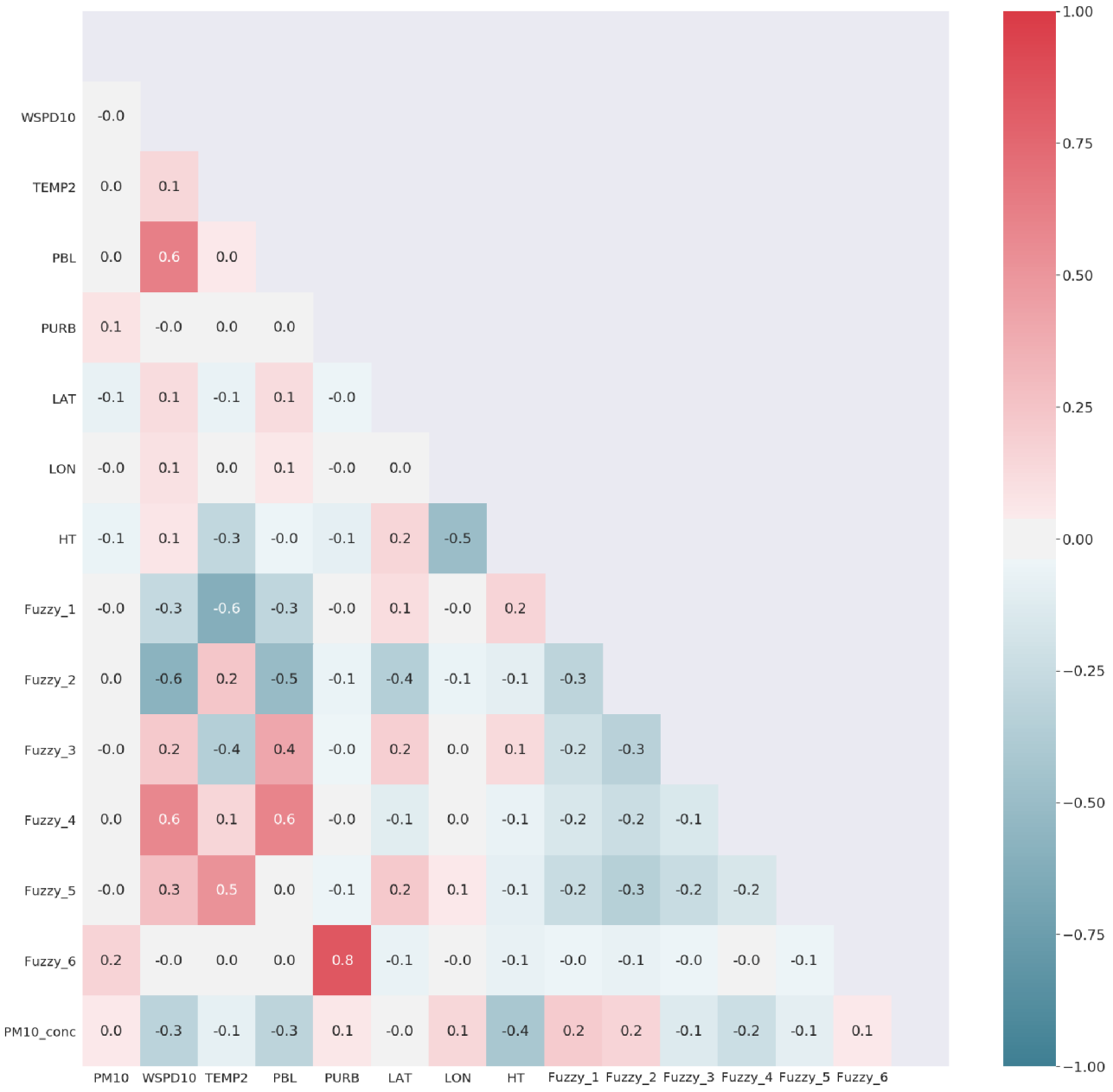
Obr. 31: Korelačne matica pre hodnoty korelačného Pearsonovho koeficientu pre dátovu maticu s pridaním matice príslušnosti k-means zhlukovania pre PM_{10}

Tabuľka 10: Mohutnosti výsledných zhlukov algoritmu fuzzy c-means s hodnotou váhového koeficientu $m = 1,5$

Zhluk	Mohutnosť zhluku
fuzzy_1	1458
fuzzy_2	875
fuzzy_3	1740
fuzzy_4	2548
fuzzy_5	1742
fuzzy_6	1321



Obr. 32: Korelačne matica pre hodnoty Pearsonovho korelačného koeficientu pre dátovu maticu s pridaním matice príslušnosti z fuzzy c-means algoritmu s $m = 1,5$



Obr. 33: Korelačne matica pre hodnoty Pearsonovho korelačného koeficientu pre dátovu maticu s pridaním matice príslušnosti z fuzzy c-means algoritmu s $m = 1,05$

Tabuľka 11: Výsledné koeficienty determinácie kvadratického regresného modelu pre výsledné zhluky k-means algoritmu

Zhluk	r^2	r^2_{adj}	RSS
kmeans_1	0,386	0,377	0,803
kmeans_2	0,488	0,478	0,577
kmeans_3	1	nan	0
kmeans_4	0,662	0,655	0,656
kmeans_5	0,487	0,163	0,746
kmeans_6	0,792	0,787	0,326
kmeans	0,585		0,644

Tabuľka 12: Výsledné koeficienty determinácie kvadratického regresného modelu pre výsledné zhluky fuzzy c-means algoritmu pre váhový exponent $m = 1.5$

Zhluk	r^2	r^2_{adj}	RSS
fuzzy_1	0,516	0,504	0,335
fuzzy_2	0,379	0,368	0,818
fuzzy_3	0,861	0,853	0,614
fuzzy_4	0,651	0,640	0,528
fuzzy_5	0,778	0,770	0,689
fuzzy_6	0,639	0,629	0,228
fuzzy$_{m=1,05}$	0,613		0,622

Tabuľka 13: Výsledné koeficienty determinácie kvadratického regresného modelu pre výsledné zhluky fuzzy c-means algoritmu pre váhový exponent $m = 1.05$

Zhluk	r^2	r^2_{adj}	RSS
fuzzy_1	0,525	0,515	0,682
fuzzy_2	0,817	0,809	0,813
fuzzy_3	0,410	0,076	0,452
fuzzy_4	0,680	0,672	0,260
fuzzy_5	0,313	0,364	0,551
fuzzy_6	0,665	0,656	0,814
fuzzy$_{m=1,05}$	0,592		0,639